Estimations précises de grandes déviations et applications à la statistique des séquences biologiques

Pierre Pudlo

Université Claude Bernard – Lyon 1 **École doctorale MathIF**

Estimations précises de grandes déviations et applications à la statistique des séquences biologiques

Thèse

Numéro d'ordre : 253 – 2004

présentée et soutenue publiquement le 16 décembre 2004

pour l'obtention du

Diplôme de Doctorat

(arrêté du 30 mars 1992)

Spécialité : Mathématiques

par

Pierre Pudlo

Composition du jury

M. Bernard Bercu	Professeur (Université Toulouse 3), Rapporteur
M. André Goldman	Professeur (Université Lyon 1)
M. Didier Piau	Professeur (Université Lyon 1), Directeur de thèse
M. Bernard Prum	Professeur (Université d'Évry), Président
M. Stéphane Robin	Professeur (Institut National Agronomique Paris-Grignon), Rapporteur

Laboratoire de Probabilité, Combinatoire et Statistique

Mis en page avec la classe thloria.

Remerciements

Je tiens tout d'abord à remercier chaleureusement Didier Piau, sous la direction duquel cette thèse a été effectuée. Sa grande disponibilité, son ouverture aux sujets de Biostatistique, ses qualités et son soutien constant ont été déterminants.

Bernard Bercu et Stéphane Robin m'ont fait un grand honneur en acceptant d'être les rapporteurs de cette thèse. Je leur exprime ma profonde gratitude pour le soin qu'ils ont porté à la lecture de mon manuscrit et pour les remarques pertinentes qu'ils ont formulées.

Je veux remercier également André Goldman pour son accueil au LaPCS lorsque j'ai commencé mon doctorat. Son soutien et son attention m'ont permis de travailler sereinement dans un laboratoire dynamique. Sa présence dans le jury m'honore particulièrement.

Je suis reconnaissant à Bernard Prum de m'avoir fait le grand honneur de présider le jury de ma thèse. Je tiens à le remercier vivement.

Je dois également remercier tous les membres du laboratoire que je cotoie depuis maintenant trois ans, qui m'ont constamment soutenu et aidé pour des questions de mathématiques, pour mes enseignements, etc. Je pense en particulier à Anne, Aurélia, Christelle, Christian, Clément, Fabien, Frédérique, Gabriela, Jean, Jean-Baptiste, Mariam, Nicolas, Pierre, Véronique et Madame Lefranc.

Je remercie également mes parents, mon frère et mes proches pour leur soutien tout au long de mes études.

Table des matières

Notations

Alphabets et mots

Soit $\mathcal{A}$ un alphabet, c'est-à-dire un ensemble de cardinal fini. Un mot de longueur ℓ est défini comme un élément de $\mathcal{A}^\ell$.

Soit x un mot sur l'alphabet $\mathcal{A}$. On note $x(i:j)$ le mot $x_i\, x_{i+1} \dots\, x_j$.

Vecteurs et matrices

Si $x,\, y \in \mathbb{R}^d$, on note $x \le y$ lorsque, pour tout j, $x_j \le y_j$. On note $x < y$ lorsque pour tout j, $x_j < y_j$. On note $x\, y$ le produit scalaire de x et y, i.e.

$$x\, y := \sum_{j=1}^{d} x_j y_j.$$

On définit aussi la partie entière inférieure et supérieure d'un vecteur x de $\mathbb{R}^d$, notée respectivement $\lfloor x \rfloor$ et $\lceil x \rceil$: $y = \lfloor x \rfloor$ est l'unique vecteur de $\mathbb{Z}^d$ dont les coordonnées vérifient $y_j \le x_j < y_j + 1$ et $z = \lceil x \rceil$ est l'unique vecteur de $\mathbb{Z}^d$ dont les coordonnées vérifient $z_j - 1 < x_j \le z_j$. De plus, si Q est une matrice de taille $d \times d$,

$$(Q\, y)_i := \sum_{j=1}^{d} Q(i,\, j) y_j, \quad 1 \le i \le d$$

$$(x\, Q)_j := \sum_{i=1}^{d} x_i Q(i,\, j), \quad 1 \le j \le d$$

$$x\, Q\, y := \sum_{i=1}^{d} \sum_{j=1}^{d} x_i Q(i, j) y_j.$$

On introduit l'ensemble

$$\mathbb{T} := [-\pi\,;\, +\pi]^d \subset \mathbb{R}^d.$$

Si Σ est un ensemble de cardinal d, les mêmes notations sont valables pour des vecteurs de $\mathbb{R}^\Sigma$ en identifiant cet espace avec $\mathbb{R}^d$.

Loi gaussienne

Soit Γ une matrice symétrique définie positive de taille $d \times d$. On note φ_Γ la densité de la loi gaussienne centrée de covariance Γ

$$\varphi_\Gamma(x) := \frac{1}{(2\pi)^{d/2}\,(\det\Gamma)^{1/2}} \exp\left(x\,\Gamma^{-1}\,x\right)$$

De plus, on note $\langle\Gamma\rangle$ la valeur à l'origine de la densité de la loi gaussienne centrée de covariance Γ :

$$\langle\Gamma\rangle := \varphi_\Gamma(0) = (2\pi)^{-d/2}\,(\det\Gamma)^{-1/2}$$

Fonctions numériques

Soit $d \geq 1$ un entier. Pour $K := (K_j)_j$ dans $\mathbb{N}^d$, et $\varphi : \mathbb{R}^d \to \mathbb{R}$ une fonction suffisamment dérivable, on note

$$z^K := \prod_{j=1}^{d} {z_j}^{K_j}, \quad D^K\varphi := \frac{\partial^{K_1+\cdots+K_d}\varphi}{\partial {t_1}^{K_1}\cdots\partial {t_d}^{K_d}}.$$

Pour tout polynôme $P(X) = \sum_K \beta_K X^K$ de $\mathbb{C}[X_1,\ldots,X_d]$ avec $X = (X_j)_j$, et les conventions ci-dessus, on définit l'opérateur différentiel $P(D)$ par

$$P(D)\varphi := \sum_K \beta_K D^K\varphi.$$

Variables et vecteurs aléatoires

Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, $d \geq 1$. On note $\mathbb{E}(X)$ l'espérance de X sur Ω et, pour tout événement A, $\mathbb{E}(X\,;\,A)$ l'espérance de X sur A, c'est-à-dire

$$\mathbb{E}(X\,;\,A) := \mathbb{E}(X\mathbf{1}_A) = \int_A X(\omega)\mathrm{d}\mathbb{P}(\omega).$$

La fonction génératrice des moments de X est définie sur $\mathbb{R}^d$ ou $\mathbb{C}^d$ par

$$I(t) := \mathbb{E}\exp(t\,X)$$

lorsque cette espérance a un sens. De même, la fonction génératrice des cumulants de X est définie sur $\mathbb{R}^d$ ou $\mathbb{C}^d$ par

$$\Lambda(t) := \log I(t) = \log \mathbb{E} \exp(t\, X).$$

Introduction

1 Motivations biologiques

Avec le séquençage, les biologistes disposent d'une grande quantité de données sur les génomes, ce qui impose l'utilisation d'outils quasi-mécaniques pour analyser ces séquences. Dans cette thèse, nous étudions des méthodes pour trouver des mots de fréquences exceptionnelles par rapport à un modèle aléatoire. Si ce modèle est bien choisi, on espère ainsi exhiber des mots dans ces séquences qui ont un rôle biologique particulier. Par exemple, chez Escherichia coli, le mot `GCTGGTGG` est sur-représenté, parce qu'il inhibe l'action des nucléases, enzymes qui détruisent l'ADN mais sont présentes dans le noyau du bacille E. coli. C'est un site chi (Crossover Hotspot Investigation). Les biologistes connaissent d'autres exemples de sites chi dans différents organismes, par exemple `CCGCT` chez Bacillus subtilis, ou tous les mots de la forme `GxTGGTGG`, $\mathtt{x} \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ chez Haemophilus influenzae. Les sites de restrictions donnent un autre exemple intéressant de mots exceptionnels. Ce sont des mots du génome à partir desquels une enzyme peut le découper, et ils sont connus pour être sous-représentés dans des modèles aléatoires, voir par exemple [49, p. 150–181]. Nous souhaitons donc faire le travail inverse, c'est-à-dire obtenir des mots ayant potentiellement un rôle biologique à l'aide d'outils statistiques.

Nous comparons donc la séquence d'ADN à une suite de variables aléatoires à valeurs dans l'alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ (de même pour une séquence d'ARN ou de protéines quitte à changer l'alphabet). On utilise des modèles issus des familles suivantes : chaînes de Markov, chaînes m-markoviennes et chaînes de Markov cachées (HMM). Nous considérons cet aléa comme le bruit sur la séquence observée : nous savons pertinemment que l'ADN n'est pas une réalisation typique de l'un de ces modèles aléatoires. Mais, ils tiennent compte d'un certain nombre de propriétés de la séquence biologique : mutations aléatoires pour les modèles m-markoviens et cartographie de la séquence en région de nature différente pour les chaînes de Markov cachées. Ainsi, en comparant ce modèle aléatoire à la séquence observée, nous espérons détecter des signaux biologiques. Plus précisément, nous allons chercher des mots sur-représentés ou sous-représentés dans la séquence observée par rapport au modèle aléatoire. La

mauvaise qualité de la prédiction des fréquences par le modèle n'est probablement pas due au hasard, mais plutôt à des nécessités biologiques, comme pour les motifs chi ou les sites de restrictions.

Soit w un mot de longueur ℓ. On dit que le mot w est sur-représenté dans la séquence observée si la fréquence des occurrences de w, notée $\alpha(w)$, est supérieure ou égale à la fréquence moyenne $n^{-1}S_n(w)$ de w dans la séquence aléatoire de longueur n, $S_n(w)$ étant le nombre d'occurrences de w dans cette séquence. Une façon naturelle de quantifier cette sur-représentation est d'utiliser

$$\mathbb{P}(n^{-1}S_n(w) \geq \alpha(w)).$$

Plus cette probabilité est petite, plus le modèle aléatoire prédit mal la fréquence réellement observée de w. Et plus cette prédiction est mauvaise, plus nous sommes en droit de supposer que cet écart entre le modèle aléatoire et la séquence a une cause biologique. De même, un mot est sous-représenté lorsque la fréquence observée est inférieure à la fréquence moyenne prédite par le modèle. Et on quantifie cette exceptionnalité par $\mathbb{P}(n^{-1}S_n(w) \leq \alpha(w))$.

La distribution exacte de $S_n(w)$ dans les modèles markoviens a été étudiée par Robin et Daudin [59], ainsi que Régnier [52]. Malheureusement, il est difficile de calculer cette probabilité directement, quand la longueur n de la séquence est grande. Nous devons alors utiliser des approximations asymptotiques de cette probabilité quand n tend vers l'infini. Dans la bibliographie, il existe deux types d'approximations. La première méthode consiste à approcher la distribution de $S_n(w)$ par des lois gaussiennes ou de Poisson composées, voir Reinert, Schbath et Waterman [55] et Schbath [60]. Le deuxième type d'approximation est de type grandes déviations. Nous utilisons ce type d'approximation, à la suite des articles de Régnier et Szpankowski [54], de Reinert, Schbath et Waterman [55] et de la thèse de Nuel [49]. Ces résultats de grandes déviations donnent le comportement exponentiel de la queue de la distribution de $S_n(w)$.

De plus, nous voulons établir une liste ordonnée de mots de fréquences exceptionnelles, plutôt que de tester la qualité de la prédiction de la fréquence d'un mot donné. Fixons la longueur ℓ des mots auxquels nous nous intéressons. Nous construisons cette liste récursivement en commençant par chercher le mot dont la fréquence est la plus mal prédite parmi tous les mots de longueur ℓ, ce qui nous donne le premier mot w_1 de la liste. Ensuite, nous conditionnons le modèle par la fréquence observée de w_1 et nous cherchons le mot qui soit le plus exceptionnel dans ce nouveau modèle. Puis, nous conditionnons le modèle par les fréquences de ces deux premiers mots, etc. Donc, nous devons étudier la distribution d'un vecteur $S_n := (S_n(w))_{w \in L}$ de nombres d'apparitions des mots de l'ensemble fini $L = \{w_j \,;\, 1 \leq j \leq d\}$.

Cette méthode semble donner de meilleurs résultats que le classement des mots de longueur ℓ en fonction de la qualité de la prédiction de leur fréquence d'apparition dans un modèle aléatoire donné. Dans cette liste obtenue sans

conditionnements successifs, il risque d'y avoir un grand nombre de mots inutiles. En effet, si un mot w est sur-représenté pour une raison biologique, il est fort probable que des mots comme $a\,w_{1:\ell-1}$ ou $w_{2:\ell}\,a$, $a \in \mathcal{A}$, qui apparaissent juste avant ou juste après w, soient eux aussi sur-représentés. Ce sont donc des mots qui viennent s'ajouter à la liste inutilement. Par exemple, pour la liste des 30 premiers mots de longueur 8 chez Escherichia coli [49, p. 167], le troisième et le vingtième mot contiennent le même sous-mot `CTGGCGG`, ainsi que le quatrième et le vingt-sixième, qui contiennent le sous-mot `GCTGGTG`. De plus, dix-mots sur ces trente mots, dont les sept premiers, contiennent le sous-mot `GCTGG` ou son complémentaire `CCAGC`. En revanche, quand on conditionne le modèle par la fréquence observée de w, il est légitime de penser que les fréquences des mots $a\,w_{1:\ell-1}$ ou $w_{2:\ell}\,a$, $a \in \mathcal{A}$ sont mieux prédites par ce nouveau modèle. Nous obtenons donc une liste plus pertinente en faisant des conditionnements successifs. Ce qui permet de minimiser les efforts à fournir pour exploiter ces listes de mots, puisqu'il faudra chercher pour chaque mot de la liste une fonction éventuelle biologique.

Pour résumer le problème mathématique, nous nous intéressons aux grandes déviations d'une fonction vectorielle du modèle aléatoire. Les fonctionnelles d'une chaîne de Markov, ou d'une chaîne de Markov cachée de n'importe quel ordre peuvent être vues comme fonctionnelles d'une chaîne de Markov simple. En effet, les chaînes de Markov cachées peuvent être écrites comme fonction d'une chaîne de Markov, et les processus markoviens d'ordre m quelconque sont des projections de chaînes de Markov d'ordre 1. Nos résultats sur l'estimation de $\mathbb{P}(n^{-1}S_n \in n\,B)$ s'appliquent donc à tous ces modèles aléatoires.

2 Principe de grandes déviations

Soit L une liste de cardinal d de mots de longueur ℓ. Nous utilisons un principe de grandes déviations (PGD) pour quantifier l'exceptionnalité de la liste L. En effet, le PGD permet de comparer des événements dont la probabilité décroît exponentiellement en fonction de la longueur n de la séquence. Rappelons que les processus de Markov additifs, construits sur une chaîne finie, irréductible et apériodique, satisfont un PGD. La suite de vecteurs $S_n \in \mathbb{Z}^d$, égaux aux nombres d'occurrences de chacun des mots de la liste L, définit un tel processus. Nous obtenons donc des estimations, à l'échelle logarithmique, des probabilités de grandes déviations de S_n. Ainsi, si $B \subset \mathbb{R}^d$ est un produit cartésien de demi-droites, il vient

$$\mathbb{P}(n^{-1}S_n \in B) = \exp(-\lambda(B)\,n + o(n)), \tag{1}$$

où $\lambda(B)$ est un coefficient positif qui dépend de B. Par exemple, si tous les mots de la liste sont sur-représentés, on choisit $B = \{x \in \mathbb{R}^d\,;\, x_j \geq v_j)\}$, en notant $v_j = \alpha(w_j)$ la fréquence observée de w_j. De plus, le PGD montre que pour tout

$\varepsilon > 0$, la probabilité de $\{n^{-1}S_n \in \{x\,;\, \forall j,\, v_j \leq x_j \leq v_j + \varepsilon\}\}$ décroît à la même vitesse exponentielle, c'est-à-dire est de la forme $\exp(-\lambda(B)\, n + o(n))$. Ce qui veut dire que lorsque S_n/n arrive à dépasser v, S_n/n reste le plus souvent au voisinage de v. En particulier, ceci nous fait dire que $\lambda(B)$ estime la qualité de la prédiction des fréquences des mots de la liste L par le modèle aléatoire.

Le théorème central limite donne lui aussi une approximation de la loi de S_n quand n est grand. Mais son intérêt est, en quelque sorte complémentaire de celui des PGD. En effet, il montre comment $n^{-1}S_n$ se concentre autour des fréquences prédites par le modèle. Par exemple, si m est la fréquence prédite par le modèle, le théorème central limite donne des estimations de probabilités qui convergent vers une constante lorsque la taille de la séquence tend vers l'infini, et plus précisément, des probabilités du type $\mathbb{P}(n^{-1}S_n \in B_n)$, où $B_n = v + E/\sqrt{n}$, $E \subset \mathbb{R}^d$. Ici, nous voulons estimer des probabilités que nous savons faibles. Dans ce cas, le théorème central limite n'est pas un très bon outil. Au contraire, le PGD donne le comportement de la queue de la distribution, mais ne dit rien du comportement typique des fréquences aléatoires $n^{-1}S_n$: dès que le vecteur des fréquences prédites par le modèle est dans l'ensemble B, $\lambda(B)$ est nul, on obtient $\mathbb{P}(n^{-1}S_n \in B) = \exp o(n)$.

En fait, un principe de grandes déviations s'énonce comme suit. Soit $(T_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans un espace polonais E. On dit que (T_n) suit un PGD de vitesse $v(n)$ et de fonction de taux (ou d'action) I si, pour tout borélien $B \subset E$

$$\begin{aligned} \liminf v(n)^{-1} \log \mathbb{P}(T_n \in B) &\geq -\inf\{I(x)\,;\, x \in \operatorname{int} B\} \\ \limsup v(n)^{-1} \log \mathbb{P}(T_n \in B) &\leq -\inf\{I(x)\,;\, x \in \bar{B}\}, \end{aligned}$$

où $\operatorname{int} B$ et $\bar{B}$ désignent respectivement l'intérieur et l'adhérence de B pour la topologie de E. Lorsque les deux infimums sont égaux, on obtient alors la limite de $v(n)^{-1} \log \mathbb{P}(T_n \in B)$. Ainsi énoncé, le PGD est très souple. En particulier, si F est un autre espace polonais et $f : E \to F$ une fonction continue, le PGD est encore vrai pour $(f(T_n))_{n\in\mathbb{N}}$, puisque

$$\{f(T_n) \in B\} = \{T_n \in f^{-1}(B)\}$$

et la continuité de f assure le fait que l'image réciproque de l'intérieur (respectivement l'adhérence) de B est l'intérieur (respectivement l'adhérence) de l'image réciproque de B. C'est ce que l'on appelle le principe de contraction. De plus, insistons sur le fait que l'énoncé du PGD repose sur la topologie de E. Dans les cas où nous utilisons un PGD, le processus aléatoire est toujours à valeurs dans un espace vectoriel de dimension finie et la topologie que nous utilisons est toujours la topologie d'espace vectoriel normé.

En fait, nous espérons que les événements $\{S_n \in n\, B\}$ associés à des petites valeurs de la fonction de taux se produiront beaucoup plus souvent que des

événements associés à des grandes valeurs de cette fonction taux. Malheureusement, le PDG ne justifie le raisonnement précédent que pour le passage à la limite $n \to +\infty$, alors que nous l'utilisons comme un outil pour comparer des fréquences de mots sur des séquences réelles, dont la longueur est finie. Nous nous trouvons donc confrontés au problème suivant : à partir de quelle longueur n pouvons nous considérer que nous sommes dans le régime asymptotique et que la valeur de la fonction de taux associée à l'événement que nous considérons a un sens ? Pour répondre à cette question, nous devons préciser le comportement de $\exp o(n)$ qui apparaît dans l'équation (1). Donnons un exemple caricatural. Supposons que les fréquences observées de deux listes L_1 et L_2 dévient des valeurs attendues par le modèle. Supposons que la probabilité qui quantifie la qualité de la prédiction des fréquences de L_1 soit égale à $\exp(-n - 10^{13})$ et que pour L_2, elle soit égale à $\exp(-2n + 100n^{7/8} - 10^{13})$. (La constante 10^{13} nous assure que ces deux nombres sont plus petits que 1 pour tout n.) Alors, pour $n \ll 10^{16}$, c'est-à-dire pour un grand nombre de cas concrets, les fréquences observées des mots de L_1 sont beaucoup plus exceptionnelles que les fréquences observées des mots de L_2, alors que la simple comparaison des actions $\lambda(L_1) = 1$ et $\lambda(L_2) = 2$ du PGD ne conduit pas à la bonne conclusion.

3 Estimations précises de grandes déviations

Il existe de nombreux résultats d'estimations précises de grandes déviations quand S_n est une somme de variables aléatoires indépendantes et identiquement distribuées (i.i.d.). Par exemple le théorème Bahadur et Rango Rao (énoncé dans le chapitre 2, théorème 2.3) donne un équivalent de $\mathbb{P}(S_n \geq nv)$ quand S_n est une somme de variables aléatoires i.i.d à valeurs dans $\mathbb{R}$:

$$\mathbb{P}(S_n \geq nv) \sim C\, n^{-1/2} \mathrm{e}^{-n\, I(v)},$$

où I est la fonction de taux donnée par le PGD de $n^{-1}S_n$. C'est le théorème le plus simple d'estimations précises de grandes déviations. Lorsque S_n est à valeurs dans un espace vectoriel de dimension d, les estimations se compliquent et dépendent de la géométrie de B. Ainsi, Ney [44] donne le comportement asymptotique de $\mathbb{P}(S_n \in n\, B)$ quand S_n est à valeurs dans $\mathbb{R}^d$ et B est un ensemble convexe d'intérieur non vide. Il montre que le comportement de cette probabilité dépend de la forme de B au voisinage d'un point v particulier de B, qu'il appelle point dominant B. Nous savons que le PGD implique que, si B est de la forme $\{x \in \mathbb{R}^d\,;\, x \geq v\}$, les rares fois où S_n/n dépasse v, S_n/n reste plutôt au voisinage de v. Suivant la forme de B au voisinage de ce point, ceci est plus ou moins difficile à réaliser. Ce qui explique que le comportement asymptotique de $\mathbb{P}(S_n/n \in B)$ dépend de la forme de B au voisinage de v. Le point dominant B est exactement le point v de B qui permet de généraliser

le raisonnement heuristique que nous avons fait ci-dessus. C'est le point vers lequel S_n/n converge conditionnellement à $\{S_n \in n\,B\}$. Comme Iltis [30] qui a développé les résultats de Ney de façon plus systématique, Ney obtient des résultats de la forme

$$\mathbb{P}(S_n \in n\,B) \sim C\,n^{-\gamma}\mathrm{e}^{-nI(v)},$$

où v est le point dominant B et γ une constante comprise entre 0 et $d/2$ qui dépend de la géométrie de B au voisinage de v. Andriani et Baldi [2] donnent une interprétation géométrique du paramètre C qui apparaît dans cet équivalent lorsque B est régulier au voisinage de v. On peut également trouver d'autres résultats sur les sommes de variables aléatoires i.i.d. dans les articles de Borovkov et Mogul′skiĭ [11, 12, 13, 14], et dans l'article de Barbe et Broniatowski [6]. Dans ce dernier article, les auteurs obtiennent un équivalent de $\mathbb{P}(S_n \in n\,B)$ pour tous les boréliens B. Ils montrent que l'on peut déduire les résultats des précédents articles en utilisant cet équivalent.

Nous nous plaçons dans le contexte mathématique différent, nos modèles étant markoviens. On se donne une chaîne de Markov $(X_j)_{j\in\mathbb{N}}$ irréductible, apériodique, sur un espace d'état fini Σ, de noyau de transition Q et de probabilité invariante π. Nous voulons étudier le processus de Markov additif $(S_n)_{n\geq 1}$ sur $\mathbb{Z}^d$ défini par

$$S_n := \sum_{j=0}^{n-1} f(X_j),$$

où $f : \Sigma \to \mathbb{Z}^d$ est une fonction définie sur l'espace d'états de la chaîne de Markov à valeurs dans $\mathbb{Z}^d$. Pour les modèles de séquences biologiques envisagés ici, le processus de comptage d'une liste de mots fixée s'écrit comme un tel processus de Markov additif.

Les événements dont nous voulons estimer la probabilité sont de événements de grandes déviations, et sont plus précisément de la forme $\{S_n \in n\,B\}$, où B est un produit cartésien de demi-droites. Quitte à changer le signe de certaines coordonnées de f, on peut supposer que l'on s'intéresse à $\{n^{-1}S_n \geq v\}$, avec v dans $\mathbb{R}^d$. Comme $n^{-1}S_n$ est un barycentre de points de $f(\Sigma) := \{f(a)\,;\,a \in \Sigma\}$, le seul cas intéressant est celui où v est dans l'enveloppe convexe de $f(\Sigma)$ et, pour éviter les cas extrémaux, on suppose que v est dans l'intérieur de $f(\Sigma)$. De plus, pour que l'événement $\{n^{-1}S_n \geq v\}$ soit exceptionnel, on suppose que $v > \pi\,f$, c'est-à-dire que toutes les coordonnées de v sont strictement supérieures aux coordonnées de la valeur moyenne de S_n/n. Comme S_n prend ses valeurs sur le réseau $\mathbb{Z}^d$, nous pouvons écrire $\{S_n \geq n\,v\} = \{S_n \geq s_n\}$, où $s_n \in \mathbb{Z}^d$ (précisément, $s_n = \lceil nv \rceil$). En fait, c'est la trace de $n\,B$ sur le réseau $\mathbb{Z}^d$ qui joue un rôle crucial. Nous avons dit que lorsque S_n dépasse nv, S_n/n reste le plus souvent au voisinage de v. En fait, $s_n = \lceil nv \rceil$ est le point le plus proche de nv que S_n puisse atteindre en dépassant nv. Nous énonçons nos théorèmes sous une forme un peu plus générale, en s'intéressant à des

événements du type $\{S_n \geq s_n\}$ où $(s_n)_{n\in\mathbb{N}}$ est une suite à valeurs dans $\mathbb{Z}^d$ qui vérifie $s_n = nv + o(\sqrt{n})$.

Nous obtenons les résultats d'estimations précises de grandes déviations suivantes. Notons $\Lambda^\star$ la fonction de taux associée au PGD de $n^{-1}S_n$. Fixons $v > \pi(f)$ dans l'intérieur de l'enveloppe convexe de $f(\Sigma)$. Donnons nous une suite (s_n) à valeurs dans $\mathbb{Z}^d$ telle que $s_n = n\,v + o(\sqrt{n})$. Alors nous obtenons

$$\mathbb{P}_a(S_n = s_n) \sim C\, n^{-d/2} \exp\left\{-n\,\Lambda^\star(v) - t\,(s_n - n\,v)\right\},$$

où t est un vecteur constant qui dépend de v et C une constante strictement positive qui dépend de a et de v. Nous obtenons le même résultat pour $\mathbb{P}_a(S_n \geq s_n)$ avec une autre constante C', elle aussi explicite. Ces deux résultats sont énoncés dans les corollaires 6.3 et 6.5 du chapitre 6. De plus, ceci montre que, pour tout borélien B' qui contient la suite $n^{-1}s_n$ et qui est contenu dans B, le comportement de $\mathbb{P}(S_n \in n\,B')$ sera toujours du même ordre. Bref, la probabilité que S_n/n dépasse v est asymptotiquement proportionnelle à la probabilité que S_n/n reste au point le plus proche de v que cette variable aléatoire puisse atteindre.

Le fait que $\mathbb{P}_a(S_n \in n\,B)$ soit équivalent à une suite u_n signifie que, pour tout $\varepsilon > 0$, il existe un rang $n(\varepsilon)$, tel que pour tout $n \geq n(\varepsilon)$, on a

$$|\mathbb{P}_a(S_n \in n\,B) - u_n| \leq \varepsilon\,|u_n|.$$

Malheureusement, nos résultats ne nous permettent pas encore de déterminer $n(\varepsilon)$ comme une fonction simple des paramètres du modèle. Donc, nous n'avons par entièrement répondu à la question de savoir à partir de quel rang n est-on dans le régime des grandes déviations, mais nous avons un développement asymptotique du préfacteur $\exp o(n)$ de l'équation (1)

Pour démontrer ce résultat, nous faisons un changement de loi sur la chaîne de Markov, en introduisant la chaîne de Markov twistée, voir chapitre 3, définition 3.11. Nous notons $\mathbb{P}^{(t)}$ une mesure de probabilité qui fait de X la chaîne de Markov twistée et $\mathbb{E}^{(t)}$ l'espérance sous cette mesure. Nous obtenons alors une égalité, valable pour tout n, qui permet de mettre en facteur l'exponentielle décroissante $\mathrm{e}^{-n\Lambda^\star(v)}$ et de réduire le problème à l'estimation de l'espérance $\mathbb{E}^{(t)}(h_n(U_n)g(X_n))$ qui dépend de la chaîne de Markov twistée, avec $U_n = n^{-1/2}(S_n - nv)$, $h_n(u) = \exp(-\sqrt{n}t\,u)\mathbf{1}_{\{u\geq 0\}}$. Cette formule de représentation montre que

$$\mathbb{P}_a(S_n \in n\,B) = \mathrm{e}^{-n\Lambda^\star(v)}\,\mathbb{E}_a^{(t)}(h_n(U_n)g(X_n)).$$

La chaîne de Markov twistée est en fait un élément d'une famille exponentielle de chaîne de Markov, construite à partir du noyau de transition Q de la chaîne initiale, sur le modèle des processus conjugués (voir chapitre 2). Mais nous

utilisons les spécificités des chaînes de Markov, en particulier la convergence de la suite de fonctions génératrices des cumulants de S_n : pour tout $t \in \mathbb{R}^d$,

$$n^{-1}\Lambda_a(n,\, t) := n^{-1} \log \mathbb{E}_a(\mathrm{e}^{t\, S_n})$$

converge, quand n tend vers l'infini vers une fonction lisse $\Lambda(t)$. De plus $\mathrm{e}^{\Lambda(t)}$ est la valeur propre dominante de la matrice $Q(t)$ définie par

$$Q(t)(a,\, b) := Q(a,\, b) \exp(t\, f(a)).$$

Le vecteur propre à droite correspondant $G(t)$ a des coordonnées strictement positives, puisque $Q(t)$ est une matrice de Perron-Frobenius. Donc, le noyau $Q^{(t)}$ défini par

$$Q^{(t)}(a,\, b) := Q(t)(a,\, b) G(t)_b / G(t)_a$$

est un noyau de transition. C'est le noyau de transition de la chaîne de Markov twistée. De plus, si t est choisi pour que $\Lambda'(t) = v$, la dérive moyenne de S_n sous la mesure twistée $\mathbb{P}^{(t)}$ est v, i.e. S_n/n converge $\mathbb{P}^{(t)}$-p.s. vers v. Et les principales contributions de S_n dans cette espérance proviennent de valeurs de S_n voisines de nv, c'est-à-dire des valeurs de U_n proches de l'origine. Une façon de le voir est de remarquer que h_n converge uniformément vers la fonction partout nulle, sauf en 0. De plus, sous la mesure twistée $\mathbb{P}^{(t)}$, $(S_n)_{n\in\mathbb{N}}$ est encore un processus de Markov additif. Nous avons donc réduit le problème à l'estimation précise de $\mathbb{P}_a(S_n = x)$ quand x est proche de nm, $m = \pi\, f$ étant la dérive moyenne de S_n. C'est ce que nous faisons avec le développement de Edgeworth du théorème 5.4 dans le chapitre 5. La recherche du rang $n(\varepsilon)$ à partir duquel on est dans le régime des grandes déviations se réduit donc à la recherche du rang à partir duquel le développement de Edgeworth est une bonne approximation de $\mathbb{P}_a(S_n = x)$.

4 Développement de Edgeworth

Notre développement de Edgeworth donne une estimation uniforme de la densité de la loi de S_n contre la mesure de comptage de $\mathbb{Z}^d$. Autrement dit, nous obtenons un développement de $\mathbb{P}_a(S_n = x)$ pour toute valeur de $x \in \mathbb{Z}^d$ fixée. Pour cela, nous construisons explicitement une famille de fonctions bornées $(\psi_a^k\, ;\, k \geq 0,\, a \in \Sigma)$ à partir de combinaisons linéaires de dérivées partielles de la densité de la loi gaussienne de covariance Γ, les coefficients de ces combinaisons linéaires étant donnés par les développements en série entière en 0 de la valeur propre dominante $\Lambda(t)$ de $Q(t)$ et d'un vecteur propre à droite associé $G(t)$. Notons $m = \pi f$ la dérive de S_n. Nous montrons qu'il existe une suite de constantes $(C_k)_{k\geq 0}$ telles que, pour tout $k \geq 0$, tout $a \in \Sigma$ et tout $n \geq 1$,

$$\left| \mathbb{P}_a(S_n = x) - n^{-d/2} \sum_{j=0}^{k} n^{-j/2}\, \psi_a^j \left(\frac{x - nm}{\sqrt{n}} \right) \right| \leq C_k\, n^{-(k+1+d)/2}. \qquad (2)$$

Ce résultat s'apparente au développement de Edgeworth pour les sommes de variables aléatoires i.i.d., voir chapitre XVI du livre de Feller [24]. Remarquons que lorsque x est loin de $n\,m$, les deux quantités $\mathbb{P}_a(S_n = x)$ et son approximation sont de toute façon chacune négligeable devant $n^{-(k+1+d)/2}$. Ce résultat est donc intéressant uniquement dans le cas où x n'est pas très loin des valeurs typiques de S_n, i.e. lorsque x est voisin de nm.

Pour démontrer ce résultat, nous utilisons une approximation de la fonction caractéristique de S_n donnée dans le chapitre 4 :

$$\mathbb{E}_a(\mathrm{e}^{\mathrm{i}\,tS_n}) \sim G(\mathrm{i}\,t)_a \exp(n\Lambda(\mathrm{i}\,t)),$$

l'erreur étant uniformément exponentiellement plus petite que $\exp(n\Lambda(\mathrm{i}\,t))$ pour tout t dans un voisinage de l'origine. Cette approximation remplace l'écriture triviale $\mathbb{E}(\mathrm{e}^{\mathrm{i}\,tS_n}) = \mathrm{e}^{n\Lambda(\mathrm{i}\,t)}$ que l'on obtient lorsque S_n est une somme de variables aléatoires i.i.d. Avec cette approximation, nous utilisons l'inverse de la transformation de Fourier pour obtenir le développement (2). Nous utilisons aussi la décroissance exponentielle du module $\mathbb{E}_a(\mathrm{e}^{\mathrm{i}\,tS_n})$ avec n quand $t \neq 0$ pour restreindre les intégrales à évaluer au voisinage de 0.

Nous utilisons ensuite ce développement pour estimer l'espérance

$$\mathbb{E}^{(t)}(h_n(U_n)g(X_n))$$

sur la chaîne de Markov twistée qui apparaît dans la formule de représentation, avec

$$h_n(u) = \exp(-\sqrt{n}t\,u)\mathbf{1}_{\{u\geq 0\}} \quad \text{et} \quad U_n = n^{-1/2}(S_n - nv).$$

Comme le théorème central limite montre que, sous $\mathbb{P}^{(t)}$, U_n converge en loi vers une gaussienne, on pourrait utiliser la méthode de Stein pour obtenir une estimation de $\mathbb{E}^{(t)}h_n(U_n)$, en oubliant le terme $g(X_n)$. La méthode de Stein permet d'obtenir des bornes explicites de la distance entre $\mathbb{E}F(U)$ et l'intégrale de F sous une loi gaussienne, où F est une fonction numérique et U une variable aléatoire. Elle donne des bornes convenables lorsque la loi de U n'est pas très loin de la loi gaussienne. Mais la méthode de Stein s'adapte mal aux fonctions non lisses comme h_n qui ont des sauts. De plus, ici, h_n converge vers la fonction indicatrice de l'origine. On est donc très loin d'une fonction lisse et nous avons préféré utiliser une méthode basée sur l'inversion de la fonction caractéristique. À ma connaissance, les résultats comme ceux de Rinott et Rotar [56, 57] ne donnent rien sur l'espérance qui nous intéresse. Il existe de nombreux résultats de développements de Edgeworth sur $\mathbb{E}(h(U_n))$, lorsque h est lisse, voir par exemple Götze et Hipp [27], Lahiri [37].

Ce qui manque à l'approximation (2) est une majoration simple de C_k en fonction des paramètres du modèle. (Voir l'annexe sur les questions ouvertes) C'est en reportant une telle majoration dans la formule de représentation avec la chaîne de Markov twistée que l'on pourra déterminer le rang à partir duquel le régime asymptotique donné par les corollaires 6.3 et 6.5 est atteint.

5 Comparaison avec d'autres résultats

Isoce, Ney et Nummelin [31] montrent que, pour tout borélien B convexe d'intérieur non vide, il existe deux constantes positives C_1 et C_2 telles que pour tout $n \geq 1$,

$$C_1\, n^{-d/2}\, \mathrm{e}^{-n\,K} \leq \mathbb{P}_a(S_n \in nB) \leq C_2\, \mathrm{e}^{-n\,K}$$

lorsque B admet un unique point dominant v, en posant $K = \Lambda^\star(v)$. Nous utilisons les mêmes techniques, c'est-à-dire le noyau twisté de la définition 3.11 et la formule de représentation de la proposition 3.12 (voir aussi p. 383–389 de [31]). Notre théorème 6.4 précise donc le comportement de cette probabilité quand B est un produit de demi-droite. Le corollaire 6.5 de ce théorème montre même que cette probabilité est équivalente à une quantité du même ordre de grandeur que la borne inférieure de cet encadrement. En effet, nous avons montré dans le théorème 5.4 un développement plus précis de $\mathbb{P}_a(S_n = x)$ que le théorème local limite utilisé par Iscoe et al. Les hypothèses sur la chaîne de Markov sont différentes des nôtres. Sans faire d'hypothèse sur la finitude de Σ, ils supposent qu'il existe une mesure de probabilité ν sur $\Sigma \times \mathbb{R}^d$, un entier $n \geq 1$ et deux nombres strictement positifs c_1 et c_2 tels que

$$c_1\, \nu(E \times \Gamma) \leq \mathbb{P}_a(X_n \in E;\, S_n \in \Gamma) \leq c_2\, \nu(B \times \Gamma)$$

pour tout état initial $a \in \Sigma$, toute partie mesurable E de Σ et tout borélien Γ de $\mathbb{R}^d$. En particulier cette hypothèse implique, quand Σ est fini qu'il existe $b_0 \in \Sigma$ tel que $Q(a,\, b_0) > 0$ pour tout a. Cette hypothèse est malheureusement impossible à faire sur de nombreux modèles markoviens des séquences génomiques.
Ney et Nummelin [45, 46] ont amélioré ces résultats pour obtenir un résultat de grandes déviations avec les hypothèses les plus faibles possibles sur la chaîne de Markov. Dans le premier article, ils simplifient la construction de la chaîne de Markov twistée en utilisant le renouvellement de la chaîne. Ce qui leur permet d'obtenir un premier PGD sur des ensembles du type

$$\{S_n \in nB\,;\, X_n \in A\}$$

avec des conditions sur la partie A de Σ. Dans le deuxième article, ils affaiblissent les conditions sur la chaîne de Markov sous lesquelles leur PGD est vrai. Mais ils n'obtiennent pas de résultats plus précis que celui énoncé dans l'article [31].

Lorsque la chaîne de Markov est réversible, définie sur un espace d'état fini, León et Perron [38] ont montré une borne supérieure pour les processus de Markov additifs à valeurs dans $\mathbb{Z}^1$, la loi initiale de la chaîne étant la distribution stationnaire π. Ils obtiennent

$$\mathbb{P}_\pi(S_n \geq nv) \leq \mathrm{e}^{-n\,K},$$

où K est un nombre positif qui dépend de la dérive πf de S_n, des points extrémaux du support de f et de la seconde valeur propre ρ_2 du noyau de transition. Quand ρ_2 est positive, León et Perron montrent que K est en fait la fonction de taux donnée par principe de grandes déviations. Notons que nous n'avons pas besoin de la réversibilité de la chaîne, ni de l'hypothèse $\rho_2 \geq 0$ pour obtenir nos résultats. De plus, même lorsque $\rho_2 \geq 0$, nous démontrons que l'équivalent de $\mathbb{P}_\pi(S_n \geq nv)$ est $o(\mathrm{e}^{-n\,K})$, puisqu'il y a dans l'équivalent un facteur en $n^{-1/2}$ dans le cas de la dimension 1. Les bornes comme celles de León et Perron sont souvent utilisées pour justifier des tests statistiques, en donnant une majoration de la puissance du test. Ici, notre problème est différent, puisque nous devons comparer la mauvaise qualité de prédictions de fréquences de mots dans un modèle aléatoire. Nous avons donc vraiment besoin d'un encadrement ou d'un équivalent.

Dans sa thèse, Nelly Torrent [62, théorème 7.1] a donné le terme dominant du développement asymptotique donné par le théorème 6.4, c'est-à-dire, le corollaire 6.5 dans le cadre markovien unidimensionnel.

Dans un article récent, Kontoyiannis et Meyn [35] se sont intéressés aux processus de Markov additifs en dimension $d = 1$, lorsque la chaîne de Markov est définie sur un espace d'état quelconque, avec une condition d'ergodicité géométrique ([35, p. 16–32]). Ils montrent que les techniques de valeurs propres utilisées pour construire le noyau $Q(t)$, $t \in \mathbb{R}^d$ s'adaptent au cas où t est complexe, $t \in \mathbb{C}^d$. C'est une piste intéressante pour affaiblir les hypothèses du chapitre 4 de cette thèse et en déduire des estimations précises de grandes déviations dans un cadre théorique plus large. Nous avons séparé ce petit chapitre préliminaire du chapitre énonçant le développement de Edgeworth pour insister sur le fait que celui-ci ne dépend de l'hypothèse de finitude de Σ que pour appliquer les estimations sur les fonctions caractéristiques obtenues dans le chapitre 4. Donc, si les techniques de Kontoyiannis et Meyn permettent de montrer les résultats sur la fonction caractéristique du chapitre 4 sous des hypothèses plus faibles, nous pouvons en déduire directement que notre développement de Edgeworth est vrai dans ce cas de figure.

6 Quelques résultats sur le génome d'Escherichia Coli

Le premier problème pour appliquer les résultats de grandes déviations dans une situation concrète est le calcul de la fonction de taux, puis sa minimisation sur les ensembles qui nous intéressent.

Une première piste pour estimer cette fonction de taux est de calculer la valeur propre dominante $\Lambda(t)$ de $Q(t)$ et d'en déduire les valeurs de la transformée de

Fenchel-Legendre de cette fonction $\Lambda^\star$. Malheureusement, cette méthode que Nuel [49] a appliqué dans sa thèse n'est pas simple. Par contre, nous savons que le minimum de la fonction de taux qui nous intéresse est atteint en $\Lambda^\star(v)$, v étant le vecteur des fréquences observées.

Une deuxième piste, qui est celle que nous avons exploitée, est d'utiliser le fait que la fonction de taux de la mesure empirique des fréquences des mots de longueur ℓ s'écrit simplement. Puis d'utiliser le principe de contraction pour obtenir le minimum de cette fonction de taux sur le bon ensemble. La fonction de taux K coïncide avec une fonction convexe régulière J sur un ensemble $\mathcal{S}$ de mesures invariantes par translation, et est infinie partout ailleurs. Il y a deux méthodes pour calculer le minimum de cette fonction de taux sur une partie E de $\mathcal{S}$. La première méthode est un algorithme de descente du gradient en projetant le gradient pour rester dans E, c'est-à-dire pour respecter la contrainte. C'est celle que nous avons employée. Le calcul de la projection du gradient limite beaucoup l'efficacité de cette méthode. En particulier, il faut calculer une matrice de projection sur $\Sigma \times \Sigma$, ce qui est coûteux en temps et en mémoire, Σ étant l'espace d'état du modèle aléatoire des mots de longueur ℓ. Le calcul de cette projection nous limite également dans le choix du modèle. Si on choisit un modèle de type chaîne de Markov cachée, la taille de Σ augmente beaucoup et l'implémentation que nous avons faite de l'algorithme de minimisation ne tourne plus. Les calculs que nous avons menés sont donc limités à des modèles m-markoviens, $m \leq 3$ et $\ell \leq 5$. La deuxième méthode, que nous n'avons pas encore implémentée, serait de pénaliser J lorsqu'on sort de E, et appliquer un algorithme de descente du gradient classique. La partie délicate est de choisir la pénalisation pour obtenir la convergence souhaitée.

De plus, il faut que le modèle prédise suffisamment mal les fréquences des mots de longueur ℓ pour que cette méthode donne de bon résultats. Nous avons par exemple essayé de faire tourner cette méthode sur un modèle 3-markovien pour des mots de longueur 5, i.e. un modèle markovien réglé pour que les fréquences prédites des mots de longueur inférieure ou égale à 4 soient celles observées. Mais les résultats obtenus sont très décevants : les éventuels signaux que l'on veut détecter sont trop proches du bruit que modélise la séquence aléatoire.

Notons que la plupart des mots exceptionnels obtenus ont leurs symétriques complémentaires dans la liste de mots exceptionnels. Ce qui est encourageant : si ces mots sont effectivement sur-représentés ou sous-représentés sur un brin d'ADN pour des raisons de stabilité, il est fort probable qu'ils soient sur-représentés ou sous-représentés sur le brin complémentaire pour les mêmes raisons. Malheureusement, les limites sur la longueur des mots ne permettent pas de faire apparaître des motifs connus, ceux-ci étant de longueur plus grande que 5.

Chapitre 1

Principe de grandes déviations

Sommaire

Un principe de grandes déviations (PGD) donne la décroissance exponentielle de la queue de la distribution en fonction de n d'une suite de variables aléatoires $(X_n)_n$. Ce chapitre présente rapidement les principes de grandes déviations et quelques outils auxquels nous faisons référence dans la suite de cette thèse. Les définitions et les résultats présentés ici sont issus du livre de Dembo et Zeitouni [21].

Commençons par regarder un exemple simple. Soit (S_n) une suite de variables aléatoires, définie pour tout $n \geq 1$, par

$$S_n = \sum_{k=0}^{n-1} Y_k,$$

où (Y_k) est une suite de variables aléatoires i.i.d. réelles. On suppose que les Y_k sont centrées et que, pour tout $t \in \mathbb{R}$, $\exp tY_0$ est intégrable et on note Λ la fonction génératrice des cumulants de Y_0. On note

$$\Lambda^\star(x) = \sup\{tx - \Lambda(t)\,;\, t \in \mathbb{R}\}.$$

Soit $x \geq 0$. L'inégalité de Markov montre que, pout tout $t \geq 0$,

$$\mathbb{P}(S_n - nx \geq 0) \leq \mathbb{E}(\exp(t\,(S_n - nx))) = \mathrm{e}^{-n(tx-\Lambda(t))}, \tag{1.1}$$

car $\inf\{\mathrm{e}^{t\,y}\,;\, y \geq 0\} = 1$. On peut alors choisir le meilleur de ces majorants en posant

$$K(x) := \sup\{tx - \Lambda(t)\,;\, t \geq 0\}.$$

De plus, pour $x \geq 0$, $K(x) = \Lambda^\star(x)$ car les variables aléatoires Y_k sont centrées. En effet, l'inégalité de Jensen montre que, pour tout réel t, $\Lambda(t) \geq 0$, donc $\Lambda^\star(0) = 0$. De plus, pour tout $t < 0$, $tx - \Lambda(t) \leq -\Lambda(t)$, et par définition de $\Lambda^\star(0)$, $-\Lambda(t) \leq \Lambda^\star(0) = 0$. D'où, pour tout $t < 0$, $tx - \Lambda(t) \leq 0$.

Avec l'équation (1.1), il vient $\mathbb{P}(S_n \geq nx) \leq \mathrm{e}^{-n\Lambda^\star(x)}$ pour tout $x \geq 0$. Le principe de grandes déviations montre que ce majorant est optimal, c'est-à-dire

$$\lim \frac{1}{n} \log \mathbb{P}(S_n \geq nx) = -\Lambda^\star(x).$$

En changeant les Y_k en $-Y_k$, on obtient, pour tout $x \leq 0$,

$$\mathbb{P}(S_n \leq nx) \leq \mathrm{e}^{-n\Lambda^\star(x)}.$$

Plus précisément, le principe de grandes déviations s'intéresse aux événements de la forme $\mathbb{P}(n^{-1}S_n \in B)$ pour tous les boréliens B. En particulier, le fait d'avoir un résultat pour tous les boréliens B permet d'en déduire des encadrements, à l'échelle logarithmique de $\mathbb{P}(f(n^{-1}S_n) \in B')$ pour des boréliens B', lorsque $f : \mathbb{R} \to \mathbb{R}$ est une fonction continue. C'est le principe de contraction énoncé dans le théorème 1.7.

1.1 Définition

Soit E un espace polonais et $\mathcal{B}$ la tribu des boréliens sur E, éventuellement complétée. Soit $(\mu_n)_{n\in\mathbb{N}}$ une suite de mesures de probabilités sur E.

Définition 1.1. *Une fonction de taux I est une application $I : E \to [0; +\infty]$ semi-continue inférieurement (s.c.i), c'est-à-dire dont les ensembles de niveau $\{x \in E\,;\, I(x) \leq \alpha\}$, pour $\alpha \in \mathbb{R}_+$, sont des parties fermées de E. Lorsque les ensembles de niveau sont compacts, on dit que I est une bonne fonction de taux.*

Le PGD s'énonce comme suit. Insistons sur le fait que cette définition est liée à la topologie de E.

Définition 1.2. *La suite $(\mu_n)_{n\in\mathbb{N}}$ suit un principe de grandes déviations de vitesse $v(n)$ et de fonction de taux I si, pour tout $A \in \mathcal{B}$,*

$$-\inf\{I(x)\,;\, x \in \operatorname{int} A\} \leq \liminf_{n\to+\infty} v(n)^{-1} \log \mu_n(A)$$
$$\limsup_{n\to+\infty} v(n)^{-1} \log \mu_n(A) \leq -\inf\{I(x)\,;\, x \in \bar{A}\}.$$

Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ à valeurs dans E suit un PGD si la suite des mesures images suit un PGD.

Lorsque la vitesse n'est pas précisée, $v(n) = n$. De plus, les bonnes fonctions de taux ont certaines propriétés de régularité, voir par exemple Dembo et Zeitouni [21, p. 119].

Proposition 1.3. *Soit I une bonne fonction de taux sur un espace polonais E. Si $(F_\delta)_{\delta>0}$ est une famille croissante de fermés de E et si $F_0 = \cap_{\delta>0} F_\delta$, alors*

$$\inf\{I(x)\,;\, x \in F_0\} = \lim_{\delta\to 0} \inf\{I(x)\,;\, x \in F_\delta\}.$$

Si, de plus E est un espace métrique, et A est une partie de E, alors

$$\inf\{I(x)\,;\, x \in \bar{A}\} = \lim_{\delta\to 0} \inf\{I(x)\,;\, x \in A_\delta\},$$

où $A_\delta = \{x \in E\,;\, d(x,\, A) \leq \delta\}$.

Si $\mathcal{A}$ est une base de la topologie de E, on peut caractériser la valeur de fonction de taux en un point x à l'aide de limites sur $\mathbb{P}(X_n \in A)$, $x \in A \in \mathcal{A}$.

Proposition 1.4. *Supposons que (X_n) suit un PGD de vitesse $v(n)$ et de fonction de taux I. Alors, pour tout $x \in E$,*

$$I(x) = \sup\left\{-\liminf v(n)^{-1} \log \mathbb{P}(X_n \in A)\,;\, A \in \mathcal{A},\, x \in A\right\} \tag{1.2}$$

$$= \sup\left\{-\limsup v(n)^{-1} \log \mathbb{P}(X_n \in A)\,;\, A \in \mathcal{A},\, x \in A\right\}. \tag{1.3}$$

Cette proposition admet une réciproque. Si l'équation (1.2) définit pour tout $x \in E$ une fonction de taux I et si I vérifie (1.3), alors on obtient un principe de grandes déviations faible pour (X_n) (autrement dit, la borne supérieure du principe de grandes déviations est établie uniquement pour les ensembles A dont l'adhérence est incluse dans le complémentaire d'un ensemble de niveau $\{x \in E\,;\, I(x) \leq \alpha\}$).

Soit F une partie fermée de E. Supposons que pour tout $n \in \mathbb{N}$, X_n est presque sûrement dans F et que (X_n) statisfait un PGD, alors la fonction de taux est infinie en dehors de F. On retrouve une idée voisine de celle-ci au chapitre 8, dans le théorème 8.7, où la fonction de taux associée au PGD de la mesure empirique des mots de longueur ℓ sur un modèle markovien est infinie en dehors d'un ensemble fermé.

Le PGD permet également de déduire une loi faible des grands nombres. Supposons que $(X_n)_n$ suit un principe de grandes déviations de bonne fonction de I. Supposons de plus que I admet un minimum global, atteint en un seul point $x_{\min}$ de E. Alors, X_n converge en probabilité vers $x_{\min}$.

1.2 Quelques outils

Il existe plusieurs méthodes pour montrer qu'une suite de variables aléatoires suit un principe de grandes déviations. Donnons deux méthodes qui nous seront utiles par la suite. La première méthode est le théorème de Gärtner-Ellis, qui relie PGD et comportement asymptotique de la suite des fonctions génératrices des cumulants. La deuxième méthode est le principe de contraction, qui permet de déduire des PGD pour les images continues d'un processus satisfaisant un PGD. Avant d'énoncer le théorème de Gärtner-Ellis, rappelons la définition d'une fonction essentiellement lisse.

Définition 1.5. *Soit $I : \mathbb{R}^d \to]-\infty\,;\,+\infty]$ une fonction convexe et $\mathcal{D} = \{x \in \mathbb{R}^d\,;\, I(x) < +\infty\}$. La fonction I est dite essentiellement lisse si (i) l'intérieur de $\mathcal{D}$ est non vide, (ii) si I est différentiable sur l'intérieur de $\mathcal{D}$ et (iii) si pour toute suite (x_n) dans l'intérieur de $\mathcal{D}$ convergeant vers un point de la frontière de cet intérieur, la norme de $I'(x_n)$ tend vers $+\infty$.*

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$. On munit $\mathbb{R}^d$ de sa topologie d'espace vectoriel normé, et de la tribu $\mathcal{B}$ des boréliens complétée. Notons

$$\Lambda_n(t) := \log \mathbb{E}(\mathrm{e}^{tX_n}), \quad t \in \mathbb{R}^d,$$

la suite des fonctions génératrices des cumulants. Supposons que, pour tout $t \in \mathbb{R}^d$, $\Lambda_n(t)/n$ converge (éventuellement vers $+\infty$) vers $\Lambda(t)$ et l'origine est dans l'intérieur de $\{t \in \mathbb{R}^d\,;\, \Lambda(t) < +\infty\}$. Supposons que Λ est semi-continue inférieurement (s.c.i.) et essentiellement lisse. Notons $\Lambda^\star$ la transformée de Fenchel-Legendre de Λ, c'est-à-dire la fonction définie pour tout $x \in \mathbb{R}^d$ par

$$\Lambda^\star(x) = \sup\{t\,x - \Lambda(t)\,;\, t \in \mathbb{R}^d\}.$$

Dans ce cas, le théorème de Gärtner-Ellis montre un PGD.

Théorème 1.6. *Sous ces hypothèses, la suite de vecteurs aléatoires $(X_n/n)_n$ suit un principe de grandes déviations de vitesse n et de bonne fonction de taux $\Lambda^\star$.*

La démonstration de ce résultat commence par la preuve de la borne supérieure par l'inégalité de Markov. La démonstration de la borne inférieure utilise un changement de mesure exponentiel (en introduisant un processus conjugué, comme ceux présentés dans le chapitre 2) et en utilisant la borne supérieure pour ces nouveaux processus. Pour une démonstration précise, voir par exemple Dembo et Zeitouni [21, p. 43–51].

La définition 1.2, qui donne une estimation de $\log \mathbb{P}(X_n \in A)$ pour tous les boréliens, est extrèmement souple. En particulier, le principe de contraction permet d'obtenir un PGD sur $(f(X_n))_n$ quand $(X_n)_n$ suit un PGD et f est

une fonction continue. Soient E et F deux espaces polonais et $f : E \to F$ une fonction continue. Le principe de contraction donne le résultat suivant, voir par exemple Dembo et Zeitouni [21, p. 126–127] pour une démonstration de ce résultat.

Théorème 1.7. *On suppose que $(X_n)_n$ suit un PGD dans E de vitesse $v(n)$ et de bonne fonction de taux I. Alors $(f(X_n))_n$ suit un PGD dans F de même vitesse $v(n)$ et de bonne fonction de taux J, définie par*

$$J(y) := \inf\{I(x)\,;\, f(x) = y\}.$$

Lorsque l'on a un processus $X = (X_n)_{n\in\mathbb{N}}$ à valeurs dans E, on peut construire la suite des mesures empiriques $(\chi_n)_n$, i.e. χ_n est la distribution aléatoire définie sur E par

$$\chi_n = \frac{1}{n}\sum_{j=0}^{n-1} \delta_{X_j},$$

où δ_x est la mesure de Dirac en x. On peut alors se demander si (χ_n) suit un PGD. C'est ce que l'on appelle classiquement des grandes déviations de niveau 2. Par exemple, les méthodes que nous utilisons dans le chapitre 8 sont basées sur le principe de contraction et un principe de grandes déviations de niveau 2 pour les chaînes de Markov finies, irréductibles et apériodiques. Ensuite, en appliquant le principe de contraction, nous obtenons un PGD pour tous les processus de Markov additifs dont la chaîne sous-jacente est X. Nous disposons ainsi de deux formulations pour la fonction de taux associée au PGD de ces processus. La première est construite à partir de la valeur propre dominante d'un certain noyau, voir théorème 3.9. La seconde formulation est issue du principe de projection appliqué au PGD de niveau 2, voir la partie sur la mesure empirique des mots de longueur ℓ dans le chapitre 8.

Chapitre 2

Processus conjugués

Sommaire

Dans ce chapitre, nous faisons une rapide présentation des processus conjugués. Il s'agit d'une famille exponentielle de lois, que l'on appelle aussi transformation de Esscher [23]. Cette famille permet de faire des changements de lois, en particulier nous pouvons changer l'espérance de la variable aléatoire. Ce qui permet d'appliquer la méthode dite du point selle, pour obtenir des estimations de probabilités de grandes déviations par exemple. À cette occasion, nous rappelons la démonstration du théorème d'estimation précise de grandes déviations de Bahadur et Ranga Rao [3], qui est sans doute l'application la plus simple de cette méthode. Puis nous présentons l'estimation précise de grandes déviations de Chaganty et Sethuraman [17, 18], elle aussi basée sur la méthode du point selle. Nous concluons ce chapitre avec une discussion sur l'application de ce résultat aux processus de Markov additifs, qui donne des résultats moins précis que ceux que nous obtenons par la suite.

2.1 Marches aléatoires conjuguées

Définition 2.1. *Une famille de loi $(\nu^{(t)})_{t\in\Theta}$ sur $\mathbb{R}^d$ est dite conjuguée (i) si ses lois sont deux à deux absolument continues entre elles, (ii) si pour tout s, $t \in \Theta$, la densité de $\nu^{(t)}$ contre $\nu^{(s)}$ s'écrit*

$$\frac{\mathrm{d}\nu^{(t)}}{\mathrm{d}\nu^{(s}}(x) = \exp[(t-s)\,x - c(s,\,t)] \tag{2.1}$$

*et **(iii)** s'il existe $s \in \Theta$ tel que l'ensemble des paramètres Θ contient tous les $t \in \mathbb{R}^d$ pour lesquels l'équation* (2.1) *définit une distribution, $c(s,\, t)$ étant bien choisi.*

En fait, $c(s,\, t)$ est donné par la fonction génératrice des cumulants Λ_s de la loi $\nu^{(s)}$:

$$c(s,\, t) = \Lambda_s(t-s) = \log \int_{\mathbb{R}^d} \mathrm{e}^{(t-s)\,x} \nu^{(s)}(\mathrm{d}x). \tag{2.2}$$

Nous voulons maintenant regarder une famille de marches aléatoires conjuguées. Soit $(X_j)_{j\in\mathbb{N}}$ une suite de variables aléatoires i.i.d. de loi $\nu^{(t)}$ sous la mesure de probabilité $\mathbb{P}^{(t)}$. Notons $\mathbb{E}^{(t)}$ l'espérance sous $\mathbb{P}^{(t)}$ et $\nu_n^{(t)}$ la loi d'un n-uple $(X_0, \ldots,\, X_{n-1})$ sous $\mathbb{P}^{(t)}$. Alors, la loi de $\nu_n^{(s)}$ est absolument continue contre la loi de $\nu_n^{(t)}$ et

$$\frac{\mathrm{d}\nu_n^{(s)}}{\mathrm{d}\nu_n^{(t)}}(x_0, \ldots,\, x_{n-1}) = \exp[(s-t)\, s_n - n\Lambda_t(s-t)],$$

où $s_n = x_0 + \ldots + x_{n-1}$. En particulier, si E est un événement de la tribu $\mathcal{F}_n = \sigma(X_0, \ldots,\, X_{n-1})$, nous obtenons

$$\mathbb{P}^{(s)}(E) = \mathbb{E}^{(t)}\{\exp[(s-t)\, S_n - n\Lambda(s-t)]\,;\, E\}.$$

Nous pouvons étendre ce résultat par des méthodes standard à des variables aléatoires Z, $\mathcal{F}_n$-mesurables, intégrables et à valeurs dans $\mathbb{R}^d$. Il vient

$$\mathbb{E}^{(s)} Z = \mathbb{E}^{(t)}\{\exp[(s-t)\, S_n - n\Lambda(s-t)] Z\}. \tag{2.3}$$

Soit τ un temps d'arrêt pour la filtration $(\mathcal{F}_n)$. On définit la tribu $\mathcal{F}_\tau$ par

$$\mathcal{F}_\tau = \{E \subset \Omega\,;\, E \cap \{\tau \le n\} \in \mathcal{F}_n\}.$$

Si Z est une variable aléatoire intégrable et $\mathcal{F}_\tau$-mesurable, nous obtenons alors

$$\mathbb{E}^{(s)} Z = \mathbb{E}^{(t)}\{\exp[(s-t)\, S_\tau - \tau\Lambda(s-t)] Z\}.$$

Dans les applications, on ne se donne pas une famille conjuguée de loi, mais plutôt une loi ν. Nous pouvons alors choisir s arbitrairement et utiliser les équations (2.1) et (2.2) pour construire cette famille. C'est ce que l'on appelle aussi la transformation de Esscher [23].

Exemple 2.1. Soit ν une distribution sur $\mathbb{R}^d$ pour laquelle la fonction génératrice des cumulants

$$\Lambda(t) = \log \int \mathrm{e}^{t\,x} \nu(\mathrm{d}x)$$

est définie sur un domaine $\mathcal{D}$ de $\mathbb{R}^d$. Alors Λ est C^∞ dans l'intérieur de ce domaine et

$$\nu^{(t)}(\mathrm{d}x) := \exp(t\,x - \Lambda(t))\nu(\mathrm{d}x), \quad t \in \mathcal{D}$$

définit une famille de distributions conjuguées (ici $s = 0$). En faisant varier le paramètre t dans l'intérieur de $\mathcal{D}$, nous faisons varier l'espérance de X, variable aléatoire de loi $\nu^{(t)}$ sous la mesure de probabilité $\mathbb{P}^{(t)}$. En effet, avec (2.3), il vient pour tout t dans l'intérieur de $\mathcal{D}$,

$$\mathbb{E}^{(t)}(X) = \Lambda'(t).$$

2.2 Méthode du point selle

Soit $g_n : \mathbb{R}^d \to \mathbb{R}^k$ une suite de fonctions fixée. Nous voulons étudier le comportement asymptotique de $\mathbb{E}\, g_n(S_n)$, où S_n est une marche au hasard de dérive m. Si la principale contribution de $\mathbb{E}\, g_n(S_n)$ provient de valeurs de S_n d'ordre $nm + O(n^{-1/2})$, nous pouvons utiliser le théorème central limite, ou une forme locale de convergence vers une loi normale. Cependant, lorsque nous regardons des résultats de grandes déviations, par exemple $g_n(x) = \mathbf{1}_{\{x > s_n\}}$, la méthode indiquée ci-dessus ne donne pas de résultats intéressants si s_n est trop grand.

Supposons maintenant que S_n soit une marche au hasard issue d'une famille de marches au hasard conjuguées $\mathbb{E}\, g_n(S_n) = \mathbb{E}^{(t_0)} g_n(S_n)$. La méthode du point selle (saddle point method) fournit une alternative intéressante. Elle consiste à choisir le membre t_n de la famille de lois conjuguées satisfaisant

$$\mathbb{E}^{(t_n)} S_n = s_n.$$

Alors, la fonction génératrice des cumulants de $S_n - s_n$ sous $\mathbb{P}^{(t_n)}$ a un développement de Taylor au voisinage de l'origine dont le premier terme non nul est de degré 2. En négligeant le reste, nous pouvons espérer obtenir une approximation gaussienne de $S_n - s_n$. De plus, nous pouvons écrire $\mathbb{P}(S_n \geq s_n) = \mathbb{P}^{(t_0)}(S_n \geq s_n)$ comme

$$\exp\{(t_0 - t_n)\, s_n - n\Lambda(t_0 - t_n)\}\mathbb{E}^{(t_n)}[\mathrm{e}^{(t_0 - t)\,(S_n - s_n)}\,;\, S_n - s_n > 0]. \tag{2.4}$$

et utiliser ensuite un développement local limite pour obtenir une estimation de l'espérance dans l'équation ci-dessus. Ainsi, Bahadur et Ranga Rao [3] ont montré un résultat de grandes déviations précises pour des marches aléatoires sur $\mathbb{R}^1$. Nous présentons ici une forme faible de leur théorème, qui donne le terme dominant du développement que Bahadur et Ranga Rao ont montré. Notons φ la densité de la loi normale centrée réduite et rappelons un résultat local limite sur les sommes de variables aléatoires i.i.d. (voir, par exemple [24, chapitre XVI]).

Proposition 2.2. *Soit $(X_j)_{j\in\mathbb{N}}$ une suite de v.a.i.i.d. L^3 et $S_n = \sum_{j=0}^{n-1} X_j$. On note σ^2 la variance de X_j et M_3 le troisième moment de $X_j - \mathbb{E}X_j$. Alors, la densité f_n de $(\sigma^2\, n)^{-1/2}(S_n - \mathbb{E}S_n)$ vérifie*

$$f_n(x) = \varphi(x) - \frac{M_3}{6\,\sigma^3\sqrt{n}}(x^3 - 3x)\varphi(x) + o(n^{-1/2}),$$

uniformément en x quand n tend vers $+\infty$.

Théorème 2.3 (Bahadur et Ranga Rao). *Soit (X_i) une suite de variables aléatoires i.i.d. de loi ν sur $\mathbb{R}^1$. On suppose que ν est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^1$. Notons $S_n = \sum_{j=0}^{n-1} X_j$ et Λ la fonction génératrice des cumulants de X_0. Soit $t > 0$ dans l'intérieur du domaine $\mathcal{D}$ de finitude de Λ, fonction génératrice des cumulants associée à ν. On note $v = \Lambda'(t)$. Alors,*

$$\mathbb{P}(S_n \geq nv) \sim \frac{\mathrm{e}^{-n(tv-\Lambda(t))}}{t\sqrt{\Lambda''(t)2\pi n}}.$$

Démonstration. Comme $\mathbb{E}^{(t)} S_n = nv$, on applique la méthode du point selle avec $t_n = t$. L'équation (2.4) devient, avec $t_0 = 0$,

$$\mathbb{P}(S_n \geq nv) = \mathrm{e}^{-n(tv-\Lambda(t))}\mathbb{E}^{(t)}(\mathrm{e}^{-t(S_n-nv)}\,;\, S_n \geq nv) =: \mathrm{e}^{-n(tv-\Lambda(t))} I_n. \qquad (2.5)$$

Appliquons maintenant le développement de la proposition 2.2 à la densité f_n de $(\Lambda''(t)\,n)^{-1/2}(S_n - nv)$ sous $\mathbb{P}^{(t)}$: il existe $C > 0$ tel que, pour tout entier n

$$\|f_n - \varphi\|_\infty = \sup\{|f_n(x) - \varphi(x)|\,;\, x \in \mathbb{R}\} \leq C\,n^{-1/2}.$$

L'espérance I_n dans l'équation (2.5) devient, avec un changement de variables,

$$I_n = \int_0^{+\infty} \mathrm{e}^{-\sqrt{n\Lambda''(t)}\,(t\,v)} f_n(v)\mathrm{d}v = \frac{1}{\sqrt{n\Lambda''(t)}} \int_0^{+\infty} \mathrm{e}^{-t\,s} f_n\left(\frac{s}{\sqrt{n\Lambda''(t)}}\right)\mathrm{d}s.$$

Sur cette dernière intégrale appliquons le théorème de convergence dominée. La suite de fonctions $g_n : s \mapsto \mathrm{e}^{-t\,s} f_n(s/\sqrt{\Lambda''(t)\,n})$ sur $\mathbb{R}_+$ converge simplement vers $s \mapsto \mathrm{e}^{-t\,s}\varphi(0)$. De plus, pour tout entier n et pour tout s réel positif,

$$\begin{aligned}\left|f_n\left(\frac{s}{\sqrt{n\Lambda''(t)}}\right)\right| &\leq \left|(f_n-\varphi)\left(\frac{s}{\sqrt{n\Lambda''(t)}}\right)\right| + \left|\varphi\left(\frac{s}{\sqrt{n\Lambda''(t)}}\right)\right| \\ &\leq C\,n^{-1/2} + 1 \leq M.\end{aligned}$$

Donc, la suite de fonctions g_n est dominée par $s \mapsto M\mathrm{e}^{-t\,s}$ qui est intégrable sur $\mathbb{R}_+$. Il vient alors

$$\lim_{n\to\infty} \int_0^{+\infty} \mathrm{e}^{-t\,s} f_n\left(\frac{s}{\sqrt{n\Lambda''(t)}}\right)\mathrm{d}s = \int_0^{+\infty} \mathrm{e}^{-t\,s}\varphi(0)\,\mathrm{d}s = \frac{1}{t\sqrt{2\pi}}. \qquad (2.6)$$

□

Remarque 2.1. Le théorème 2.3 porte sur des variables aléatoires dont les lois sont absolument continues par rapport à la mesure de Lebesgue sur $\mathbb{R}^1$. Dans ce cas, la démonstration permet de voir que la constante $1/t$ dans l'équivalent

provient de la valeur de $\int_0^{+\infty} \mathrm{e}^{-t\,s}\,\mathrm{d}s$ (voir l'équation (2.6) dans la démonstration ci-dessus). En revanche, si le support des variables aléatoires est $\mathbb{Z}^1$, cette intégrale doit être remplacée par la série

$$\sum_{s\in\mathbb{N}} \mathrm{e}^{-t\,s} = (1-\mathrm{e}^{-t})^{-1}.$$

Et le résultat correspondant au thèorème 2.3 fait apparaître $(1-\mathrm{e}^{-t})^{-1}$ à la place de $1/t$ dans l'équivalent.

Remarque 2.2. Notons que la méthode du point selle permet aussi d'étudier une famille conjuguée de lois au bord de l'ensemble de paramètres Θ. Ainsi, Balkema, Klüppelberg et Resnick [4, 5] étudient les distributions limites d'une famille conjuguée de lois au bord de l'ensemble de paramètres Θ. Ils obtiennent des résultats de convergence faible des loi de cette famille vers des lois gaussiennes ou des loi Gamma. Barndorff-Nielsen et Klüppelberg [7] démontrent que ceci s'adapte également au cas de lois sur $\mathbb{R}^d$.

2.3 Variables aléatoires quelconques

Chaganty et Sethuraman [17] ont étudié les conditions dans lesquelles on peut adapter cette méthode à une suite quelconque de variables aléatoires $(S_n)_{n\in\mathbb{N}}$, $S_n \in \mathbb{Z}^1$. Ils obtiennent un équivalent de $\mathbb{P}(S_n \geq s_n)$ dans le régime $s_n = O(n)$.

Commençons par faire une remarque très simple dans le cas des marches aléatoires. Si (X_j) est une suite de variables aléatoires indépendantes et identiquement distribuées, de loi ν vérifiant les hypothèses de l'exemple 2.1, alors la fonction génératrice des cumulants de S_n est définie sur le même domaine $\mathcal{D}$ et vaut $n\Lambda(t)$. Cependant, dès que l'on quitte le cadre des sommes de variables aléatoires i.i.d, ceci n'est plus vrai. Dans le cas général, il faut donc introduire, pour tout $n \in \mathbb{N}$,

$$\Lambda(n,\,t) := \log \mathbb{E}\exp(t\,S_n), \quad t \in \mathbb{C}.$$

Nous notons Λ', Λ'', ... les dérivées successives par rapport à t de cette fonction. Fixons une suite $(s_n)_{n\in\mathbb{N}}$ de réels telle que $s_n = O(n)$ et supposons que nous voulons étudier $\mathbb{P}(S_n \geq s_n)$. Nous obtenons une suite de variables aléatoires conjuguées en posant

$$\mathbb{P}^{(t)}(S_n \in \mathrm{d}s) = \exp(s\,t - \Lambda(n,\,t))\mathbb{P}(S_n \in \mathrm{d}s).$$

Pour appliquer la méthode du point selle, définissons, lorsque cela est possible, la suite $(t_n)_{n\in\mathbb{N}}$ par $\mathbb{E}^{(t_n)}S_n = s_n$. La formule (2.4) devient

$$\mathbb{P}(S_n \geq s_n) = \mathrm{e}^{-(s_n\,t_n - \Lambda(n,\,t_n))}\,\mathbb{E}^{(t_n)}\left[\mathrm{e}^{-t_n\,(S_n - s_n)}\,;\,S_n - s_n \geq 0\right]. \tag{2.7}$$

Ce qui pose deux problèmes. Sous quelles hypothèses le raisonnement formel ci-dessus est-il justifié ? Sous quelles hypothèses peut -on obtenir un développement local limite de S_n sous $\mathbb{P}^{(t_n)}$?

Pour justifier la construction des processus conjugués, Chaganty et Sethuraman font les hypothèses suivantes : *(i)* il existe $a > 0$ tel que les fonctions génératrices des moments, $t \mapsto I(n, t) := \mathbb{E}(\mathrm{e}^{t S_n})$ sont définies, non nulles et analytiques sur le cercle $\mathcal{D} := \{z \in \mathbb{C}\,;\, |z| < a\}$ et *(ii)* il existe un réel $a_0 \in]0\,;\, a[$ et une suite de réels $(t_n)_{n\in\mathbb{N}}$ dans $]0\,;\, a_0[$ tels que

$$\Lambda'(n, t_n) = s_n \quad \text{pour tout } n \in \mathbb{N}.$$

Ensuite, remarquons que la transformée de Fourier de la loi de S_n sous $\mathbb{P}^{(t_n)}$ s'obtient en étudiant la fonction génératrice des moments sur $t_n + \mathrm{i}\,\mathbb{R}$. Ainsi, avec les deux hypothèses suivantes, ils obtiennent la convergence en loi de $n^{-1/2}(S_n - s_n)$ vers une loi gaussienne sous $\mathbb{P}^{(t_n)}$: *(iii)* il existe une constante finie β telle que, pour tout $n \in \mathbb{N}$, $z \in \mathcal{D}$,

$$|n^{-1}\Lambda(n, z)| \leq \beta \quad \text{et}$$

(iv) il existe une constante $g > 0$ telle que, pour tout $n \in \mathbb{N}$,

$$\Lambda''(n, t_n) \geq g.$$

Et, pour obtenir un résultat qui remplace la proposition 2.2, ils supposent de plus que *(v)* il existe $\delta_0 > 0$ tel que pour tout $0 < \delta < \delta_0$,

$$\sup\left\{\left|\frac{\Lambda(n, t_n + \mathrm{i}\,t)}{\Lambda(n, t_n)}\right| \,;\, \delta < |t| \leq \pi\right\} = o(n^{-1/2}).$$

Chaganty et Sethuraman obtiennent alors le théorème suivant. (Par des techniques similaires, ils démontrent également un résultat analogue dans le cas de variables aléatoires qui ne sont pas sur un sous-réseau de $\mathbb{R}$.)

Théorème 2.4 (Chaganty et Sethuraman). *Sous ces cinq conditions,*

$$\mathbb{P}(S_n \geq s_n) \sim \frac{1}{\sqrt{2\pi n \sigma_n}} \frac{\exp(-\gamma_n)}{1 - \mathrm{e}^{-t_n}},$$

où on a posé

$$\sigma_n = \Lambda''(n, t_n) \quad \textit{et} \quad \gamma_n = s_n\, t_n - \Lambda(n, t_n).$$

2.4 Discussion sur les processus de Markov additifs

Le cas des processus de Markov additifs est étudié en détail dans les chapitres suivants. Cependant, nous pouvons essayer de retrouver notre résultat (voir

théorèmes 6.1 et 6.4) dans le cas des processus de Markov additif sur $\mathbb{Z}^1$ à partir de ce théorème de Chaganty et Sethuraman en supposant que $s_n = nv + o(n)$, $v \geq m$, où m est la dérive du processus de Markov additif.

Pour cela, il faut montrer que les cinq conditions sont satisfaites. La condition *(i)* est vraie : avec nos hypothèses, la fonction génératrice des moments est définie sur $\mathbb{R}$. La condition *(ii)* est vraie dès l'on impose $s_n = O(n)$. Rappelons que $n^{-1}\Lambda(n,\, t)$ converge vers une fonction convexe et lisse (proposition 3.4. Alors, la conditions *(iii)* est vraie puisque une fonction lisse est bornée sur les compacts. La condition *(iv)* est vraie lorsque $s_n = nv + o(n)$, $v \neq m$. Pour vérifier la condition *(v)*, il faut utiliser des propositons similaires aux propositions 4.2 et 4.5.

Ensuite, il faut étudier le comportement asymptotique de t_n et $\exp -\gamma_n$ quand n tend vers l'infini pour obtenir un résultat utilisable dans la pratique. En fait, on peut montrer, en utilisant la proposition 3.4, que si $s_n = nv + o(n)$, alors $t_n = t + o(1)$, où t est la solution de $\Lambda'(t) = v$. Ce qui suffit pour comprendre le comportement asymptotique de $1 - \exp(-t_n)$. Malheureusement, γ_n est plus compliqué à étudier : le fait que $t_n = t + o(1)$ est vrai, mais les deux égalités $\Lambda(n,\, t_n) = n\Lambda(t) + o(1)$, et $s_n\, t_n = nv\, t + o(1)$ sont grossièrement fausses. En particulier, $\gamma_n = n(tv - \Lambda(t)) + o(1)$ est faux dans le cas général. Bref, nous ne savons pas démontrer simplement, à partir du théorème de Chaganty et Sethuraman, (sans utiliser tous les résultats des chapitres suivants) le fait que

$$\exp(-\gamma_n) \sim \exp(-t\,(s_n - nv) - n\Lambda(t)), \quad \text{quand } n \to +\infty$$

lorsque $s_n \in \mathbb{Z}^1$, $s_n = nv + o(\sqrt{n})$, comme le montre le corrolaire 6.2 de notre théorème 6.1.

De plus, à notre connaissance, nous n'avons jamais vu dans la littérature, le terme $\exp(-t\,(s_n - nv))$, même dans le cas de sommes de variables aléatoires i.i.d. à valeurs dans $\mathbb{Z}^1$ ou $\mathbb{Z}^d$, $d \geq 2$. En effet, les résultats d'estimation de grandes déviations étaient démontré à l'aide d'un développement de Edgeworth sur la fonction de répartition. Malheureusement, ce théorème porte sur une régularisation de la fonction de répartition, ce qui le rend difficile à utiliser. Notre théorème 5.4, qui donne un développement asymptotique de la densité de la loi de S_n contre la mesure de comptage de $\mathbb{Z}^d$, est plus adapté dans ce contexte, et permet d'obtenir un résultat plus précis.

Remarque 2.3. Nous avons choisi de démontrer le théorème de Bahadur et Rao à partir d'un développement asymptotique de la densité (contre la mesure de Lebesgue sur $\mathbb{R}$), plutôt que sur un développement asymptotique de la fonction de répartition, pour pouvoir l'utiliser comme point de comparaison avec les démonstrations des estimations de grandes déviations du chapitre 6.

Chapitre 3

Chaînes de Markov conjuguées

Sommaire

Dans le chapitre précédent, nous avons vu comment construire et utiliser des processus conjugués. En particulier, nous avons vu que Chaganty et Sethuraman [17] ont pu, grâce à cette technique, obtenir une estimation précise de grandes déviations sans utiliser d'hypothèse d'indépendance. Nous présentons ici une autre technique pour obtenir une famille exponentielle de chaînes de Markov. Cette nouvelle chaîne de Markov, appelée chaîne de Markov twistée, se construit à l'aide de résultat asymptotique sur la fonction génératrice des cumulants. Nous pourrons ainsi, comme dans le cas des processus conjugués, choisir la dérive de S_n. Les résultats que nous présentons ici sont dus à Iscoe, Ney et Nummelin [31], Ney et Nummelin [45, 46].

Nous nous plaçons dans le contexte mathématique suivant. Nous nous donnons une chaîne de Markov $(X_i)_{i\in\mathbb{N}}$ sur un espace d'états Σ et une fonction $f : \Sigma \to \mathbb{R}^d$. Nous étudions le processus de Markov additif $(S_n)_{n\in\mathbb{N}}$ défini par

$$S_n = \sum_{j=0}^{n-1} f(X_j).$$

Nous notons Q le noyau de transition de la chaîne de Markov, que nous supposons irréductible et apériodique. Sauf dans la section 3.3, nous supposons que Σ est fini et nous notons π la distribution stationnaire de la chaîne.

Définition 3.1. *Pour tout $t \in \mathbb{R}^d$, $Q(t)$ est le noyau défini par*

$$Q(t)(a,\, b) := Q(a,\, b) \exp t\, f(a), \quad a,\, b \in \Sigma.$$

Soit $n \in \mathbb{N}$. La fonction génératrice des moments (f.g.m.) de S_n est la fonction $t \in \mathbb{R}^d \mapsto I(n,t) \in \mathbb{R}^\Sigma$ définie avec

$$I(n,\, t)_a := \mathbb{E}_a(\exp t\, S_n), \quad a \in \Sigma.$$

Et la fonction génératrice des cumulants (f.g.c.) de S_n est la fonction $t \in \mathbb{R}^d \mapsto \Lambda(n,t) \in \mathbb{R}^\Sigma$ définie avec

$$\Lambda(n,\, t)_a := \log I(n,\, t)_a.$$

Le noyau $Q(t)$ permet d'écrire la f.g.m. de S_n comme

$$I(n,\, t) = Q(t)^n\, I.$$

où $I \in \mathbb{R}^\Sigma$ est la fonction indicatrice de Σ, définie par $I_a = 1$ pour tout $a \in \Sigma$. L'étude asymptotique de la f.g.m. de S_n se réduit donc à l'étude des itérés du noyau $Q(t)$.

Définition 3.2. *Pour tout $a \in \Sigma$, nous introduisons le temps τ_a de premier retour en a, défini par*

$$\tau_a = \inf\{j \geq 1\,;\, X_j = a\}.$$

Ce temps d'arrêt est un temps de renouvellement de la chaîne de Markov. En effet, la loi du processus $(X_{j+\tau_a})_{j\in\mathbb{N}}$ est égale à la loi de $(X_j)_{j\in\mathbb{N}}$ sous $\mathbb{P}_a$.

Nous présentons d'abord les résultats sur la valeur propre dominante de $Q(t)$. Nous rappelons que ces résultats impliquent un principe de grandes déviations (PGD). En effet, le théorème de Gärtner-Ellis permet de relier les PGD au comportement asymptotique de la f.g.c. Ensuite, nous définissons la chaîne de Markov twistée qui permet de reformuler la probabilité d'un événement de grandes déviations en mettant en facteur le terme exponentiel issu du PGD.

3.1 Valeur propre dominante

Définition 3.3. *On dit que ρ est valeur propre dominante d'un noyau si ρ est une valeur propre simple de ce noyau et si toutes les autres valeurs propres de ce noyau sont de module strictement inférieur à $|\rho|$.*

Dans le cas où l'espace d'état Σ est fini, le théorème de Perron-Frobénius montre que, pour tout $t \in \mathbb{R}^d$, $Q(t)$ admet une valeur propre dominante. De plus, cette valeur propre est strictement positive. Ce qui donne le résultat suivant.

Proposition 3.4. *Pour tout $t \in \mathbb{R}^d$, $Q(t)$ admet une valeur propre dominante $\mathrm{e}^{\Lambda(t)}$. De plus, pour tout $a \in \Sigma$, $n^{-1} \log I(n,\, t)_a$ converge, quand n tend vers $+\infty$, vers $\Lambda(t)$. En particulier, cette limite ne dépend pas de a.*

Pour montrer que $\Lambda(t)$ est analytique et essentiellement lisse, Ney et Nummelin [45] caractérisent $\Lambda(t)$ de la façon suivante.

Proposition 3.5. *Soit $a \in \Sigma$ et τ_a le temps de premier retour en a. Alors, $\Lambda(t)$ est la seule racine de l'équation*

$$\mathbb{E}_a[\exp(t\, S_{\tau_a} - \tau_a\, \Lambda(t))] = 1.$$

Remarque 3.1. Dans le cas que nous étudions, où Σ est fini, nous avons une autre équation qui définit Λ implicitement avec le polynôme caractéristique de $Q(t)$. Nous utilisons le polynôme caractéristique de $Q(\mathrm{i}\, t)$ dans le chapitre 4 pour étudier le comportement de la transformée de Fourier de S_n.

Proposition 3.6. *La fonction Λ est analytique en tout point de $\mathbb{R}^d$. De plus, Λ est strictement convexe. Son gradient en tout point t de $\mathbb{R}^d$ est donné par*

$$\Lambda'(t) = \big(E_a^{(t)} \tau_a\big)^{-1}\, E_a^{(t)} S_{\tau_a}$$

et sa matrice hessienne par

$$\operatorname{Hess} \Lambda(t) = \big(E_a^{(t)} \tau_a\big)^{-1} \operatorname{cov}_a^{(t)}(S_{\tau_a} - \tau_a\, \Lambda'(t))$$

avec les notations suivantes : pour tout v.a. Y $\mathcal{F}_{\tau_a}$-mesurable,

$$E_a^{(t)} Y := \mathbb{E}_a\big[Y \exp(t\, S_{\tau_a} - \tau_a\, \Lambda(t))\big] \quad \text{et}$$
$$\operatorname{cov}_a^{(t)} Y := \mathbb{E}_a\big[Y\ {}^t Y \exp(t\, S_{\tau_a} - \tau_a\, \Lambda(t))\big].$$

Iscoe, Ney et Nummelin [31] donnent le terme suivant dans le développement asymptotique de $I(n,\, t)$. On notera que cette fois-ci, ce terme dépend de l'état initial $a \in \Sigma$ de la chaîne de Markov.

Proposition 3.7. *Pour tout $t \in \mathbb{R}^d$, il existe un vecteur propre à droite $G(t) \in \mathbb{R}^\Sigma$ pour la valeur propre dominante $\exp \Lambda(t)$ de $Q(t)$, dont les coordonnées sont strictement positives, qui vérifie*

$$G(t) = \lim_{n \to \infty} \mathrm{e}^{-n\, \Lambda(t)} I(n,\, t).$$

De plus, $G(t)$ est donné par

$$G(t)_b = \mathbb{E}_b\big[\exp(t\, S_{\tau_a} - \tau_a\, \Lambda(t))\big], \quad b \in \Sigma$$

De même, nous obtenons un vecteur propre à gauche avec

$$\pi(t)_b = \alpha\, \mathbb{E}_a \left[\sum_{n=0}^{\tau_a - 1} \exp(t\, S_n - n\Lambda(t))\, \mathbf{1}_b(X_n) \right]$$

pour tout $\alpha \neq 0$.

Dans la suite, nous notons $\pi(t)$ ce vecteur propre à gauche, renormalisé pour que $\sum_b \pi(t)_b = 1$, ce qui fait de $\pi(t)$ une distribution sur Σ.

Remarque 3.2. Dans le cas où S_n est une somme de v.a. i.i.d., le comportement de la suite des f.g.m. est donnée par l'écriture triviale

$$\mathbb{E}(\exp t\, S_n) = \exp n\, \Lambda(t), \quad \text{où } \Lambda(t) = \log \mathbb{E}(\exp t\, X_0).$$

L'idée générale que nous utilisons par la suite est d'adapter les techniques et les résultats basés sur cette écriture au cas des processus de Markov additifs avec la proposition précédente.

Exemple 3.1. Regardons la chaîne de Markov définie sur $\Sigma = \{0\,;\,1\}$ par $Q(0,\,1) = Q(1,\,0) = \epsilon$ et le processus de Markov additif défini par $f = \mathbf{1}_1$. Le polynôme caractéristique de $Q(t)$ est

$$X^2 - (1-\epsilon)(1+\mathrm{e}^t)X + \mathrm{e}^t(1-2\epsilon)$$

et la valeur propre dominante est

$$\mathrm{e}^{\Lambda(t)} = \frac{1}{2}\left[(1-\epsilon)(1+\mathrm{e}^t) + \sqrt{(1-2\epsilon)(1-\mathrm{e}^t)^2 + \epsilon^2(1+\mathrm{e}^t)^2}\right].$$

Comme $S_{\tau_1} = 1$ presque sûrement, l'équation caractéristique de $\Lambda(t)$ donnée par la proposition 3.5 pour $a = 1$ est

$$\mathrm{e}^t\, \mathbb{E}_1 \exp(-\tau_1\, \Lambda(t)) = 1, \quad \text{avec}$$
$$\mathbb{E}_1 \exp(u\, \tau_1) = \mathrm{e}^u \left((1-\epsilon) + \frac{\mathrm{e}^u \epsilon^2}{1 - \mathrm{e}^u(1-\epsilon)}\right).$$

Avant d'énoncer le PGD de $(n^{-1}S_n)_n$, rappelons la définition de la transformée de Fenchel-Legendre d'une fonction convexe.

Définition 3.8. *Si $h : \mathbb{R}^d \to [0\,;\,+\infty]$ est une fonction convexe, sa transformée $h^\star$ de Fenchel-Legendre est définie en tout point $x \in \mathbb{R}^d$ par*

$$h^\star(x) = \sup\{t\, x - h(t)\,;\, t \in \mathbb{R}^d\}.$$

La convergence de $n^{-1} \log I(n,\, t)_a$ vers $\Lambda(t)$ et le théorème de Gärtner-Ellis permettent de montrer le principe de grandes déviations (PGD) suivant.

Théorème 3.9. *Quelque soit la distribution initiale de la chaîne de Markov, la suite $(n^{-1}S_n)$ suit un PGD de vitesse n et de bonne fonction de taux $\Lambda^\star$ dans $\mathbb{R}^d$ muni de sa topologie d'espace vectoriel normé.*

Notons $\mathcal{E}$ l'enveloppe convexe de l'image $f(\Sigma)$ de f. Nous avons les propriétés suivantes.

Proposition 3.10. *Le gradient Λ' de Λ définit un homéomorphisme de $\mathbb{R}^d$ sur l'intérieur $\mathcal{E}$. On note Ξ l'homéomorphisme réciproque.*
Soit $x \in \mathbb{R}^d$. Alors,

$$\Lambda^\star(x) = \begin{cases} \Xi(x)\, x - \Lambda \circ \Xi(x) & \text{si } x \in \operatorname{int} \mathcal{E}, \\ +\infty & \text{sinon.} \end{cases}$$

De plus, $\Lambda^\star$ admet un minimum global 0 atteint en un seul point $x = \pi\, f$.

Remarque 3.3. Il suffit de remarquer que $n^{-1}S_n$ est un barycentre de points de $f(\Sigma)$ pour voir que la fonction de taux $\Lambda^\star$ associée au PGD du théorème 3.9 est infinie en dehors de $\operatorname{int} \mathcal{E}$.

3.2 Chaîne de Markov twistée

À l'aide des propriétés de la décomposition spectrale de $Q(t)$, nous pouvons maintenant définir un changement de loi pour la chaîne de Markov.

Définition 3.11. *Pour tout $t \in \mathbb{R}^d$, $Q^{(t)}$ est le noyau de transition défini pour tout a, $b \in \Sigma$ par*

$$Q^{(t)}(a,\, b) := Q(a,\, b) \exp(t\, f(a) - \Lambda(t))\, G(t)_b\, G(t)_a^{-1}.$$

On note $\mathbb{P}^{(t)}$ une mesure de probabilité qui fait de $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov de noyau de transition $Q^{(t)}$, et $\mathbb{E}^{(t)}$ l'espérance sous cette mesure de probabilité.
De plus, le noyau $Q^{(t)}$ hérite du caractère irréductible et apériodique de Q. Le vecteur propre à gauche $\pi(t)$ donne sa distribution stationnaire.

La loi de S_n sous la mesure de probabilité $\mathbb{P}_a$ s'exprime en fonction de la loi de la chaîne twistée de la façon suivante. Pour toute fonction h définie sur $\mathbb{R}^d$ et tout état initial $a \in \Sigma$ de la chaîne de Markov, on a

$$\mathbb{E}_a(h(S_n)) = G(t)_a\, \mathbb{E}_a^{(t)} \left[h(S_n) \exp(n\, \Lambda(t) - t\, S_n)\, G(t)_{X_n}^{-1} \right].$$

et réciproquement. De plus, cette formule s'étend au cas où n est un temps d'arrêt pour la filtration associée à $(S_n)_{n\in\mathbb{N}}$. En particulier, on obtient

$$\mathrm{e}^{-n\Lambda(t)} I(n,\, t)_a = G(t)_a \mathbb{E}_a^{(t)}(G(t)_{X_n}^{-1}).$$

Le membre de gauche converge vers $G(t)_a$, alors que le membre de droite converge vers $G(t)_a \pi(t)\, G(t)^{-1}$. Nous obtenons donc l'équation suivante

$$\pi(t)\, G(t)^{-1} = 1. \tag{3.1}$$

Elle permet de déterminer, parmi les vecteurs propres à droite pour la valeur propre $\exp \Lambda(t)$ celui qui vérifie la proposition 3.7.

L'intérêt de ce changement de loi de la chaîne de Markov est de pouvoir choisir la dérive. Soit x un point de l'intérieur de $\mathcal{E}$. On fixe $t = \Xi(x)$.

Proposition 3.12. *Sous $\mathbb{P}^{(t)}$, $n^{-1}S_n$ converge presque sûrement vers x. Et, pour toute fonction h, on a*

$$\mathbb{E}_a(h(S_n)) = \mathrm{e}^{-n\,\Lambda^\star(x)}\, G(t)_a\, \mathbb{E}_a^{(t)}\left[h(S_n)\,\mathrm{e}^{-t\,(S_n - n\,x)}\, G(t)^{-1}_{X_n}\right].$$

Exemple 3.2. Soit $x \in]0;1[$. Reprenons l'exemple 3.1. La chaîne de Markov twistée, pour $t = \Xi(x)$ est l'unique chaîne de Markov dont la matrice de transition est de la forme

$$\begin{pmatrix} \dfrac{1-\epsilon}{1-\epsilon+\delta\,\epsilon} & \dfrac{\delta\,\epsilon}{1-\epsilon+\delta\,\epsilon} \\ \dfrac{\epsilon}{\delta(1-\epsilon)+\epsilon} & \dfrac{\delta(1-\epsilon)}{\delta(1-\epsilon)+\epsilon} \end{pmatrix}$$

et dont la distribution stationnaire donne le poids x à 1.

Soit $(B_n)_{n\in\mathbb{N}}$ une suite de boréliens de $\mathbb{R}^d$. Cette proposition permet d'écrire $\mathbb{P}_a(S_n \in B_n)$ en mettant en facteur le terme dominant à l'échelle logarithmique. Pour cela, il suffit de bien choisir le point $x \in \mathbb{R}^d$. En particulier, quand $B_n = n\,B$, nous rappelons la définition suivante.

Définition 3.13. *Soit B un borélien de $\mathbb{R}^d$ et ∂B sa frontière. On dit que v_B est un point dominant de B si les trois conditions suivantes sont satisfaites.*

(i) $v_B \in \partial B$.

(ii) *v_B appartient à l'enveloppe convexe de l'image de f et on note $t = \Xi(v_B)$.*

(iii) *Pour tout $x \in B$, $(x - v_B)\,t \geq 0$.*

Si B est un borélien convexe de $\mathbb{R}^d$, et v_B un point dominant B, alors $\Lambda^\star(v_B) = \inf\{\Lambda^\star(x)\,;\, x \in B\}$. Et le PGD du théorème 3.9 montre que

$$n^{-1}\log \mathbb{P}_a(S_n \in n\,B) \to -\Lambda^\star(v_B)$$

quand n tend vers ∞. Nous appliquons donc la proposition 3.12 avec $x = v_B$.

Remarque 3.4. Ceci est très proche de la première étape de la méthode du point selle présentée au chapitre 2, où l'on choisissait le membre t_n de la famille de la loi conjuguée pour que $\mathbb{E}^{(t_n)}S_n = s_n$. Comme x est la dérive de S_n sous $\mathbb{P}^{(t)}$, $U_n := n^{-1/2}(S_n - nx)$ converge en loi vers gaussienne centrée, de matrice de covariance $\Lambda''(t)$. Et

$$\mathbb{P}_a(S_n \in B_n) = \mathrm{e}^{-n\,\Lambda^\star(x)}\, G(t)_a\, \mathbb{E}_a^{(t)}\Big(h_n(U_n)G(t)^{-1}_{X_n}\Big) \tag{3.2}$$

où l'on a posé

$$h_n(u) := \begin{cases} \exp(-n^{1/2}(t\,u)) & \text{si } u \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

Une convergence en loi de U_n n'est pas suffisante pour estimer les quantités en jeu dans le membre de droite de (3.2), ni $\mathbb{E}_a^{(t)} h_n(U_n)$. En fait, la fonction h_n converge vers la fonction indicatrice de $\{0\}$ quand n tend vers l'infini. Nous avons donc plutôt besoin du comportement asymptotique de U_n au voisinage de l'origine. Nous étudions ce comportement dans le chapitre 5, dont le résultat principal est énoncé dans le théorème 5.4.

Remarque 3.5. Les méthodes à la Stein seront aussi difficiles à adapter dans ce contexte. Elles permettent de majorer efficacement la distance entre $\mathbb{E}_a^{(t)} h(U_n)$ et l'intégrale de h contre la densité de la loi gaussienne vers laquelle U_n converge. Malheureusement, même en se débarrassant de X_n dans l'espérance du membre de droite de (3.2), il reste à étudier $\mathbb{E}_a^{(t)} h_n(U_n)$, h_n étant une suite de fonctions irrégulières, qui converge vers la fonction indicatrice de $\{0\}$ quand n tend vers l'infini. Ce type d'espérance rentre très mal dans le cadre des méthodes à la Stein, même si Rinott et Rotar ont montré que la méthode de Stein s'adapte au cas des fonctions non lisses, voir [56, 57].

3.3 Quelques remarques lorsque l'espace d'états est quelconque

Dans le cas où l'espace d'état de la chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ est un espace polonais quelconque, Ney and Nummelin [45] expliquent qu'il faut faire attention à la définition de Λ pour obtenir la fonction de taux du principe de grandes déviations. En particulier, lorsque le rayon spectral est infini, il faut introduire une autre notion pour obtenir des théorèmes de grandes déviations sur

$$\mathbb{P}_a(S_n \in n\,B\,;\, X_n \in A).$$

Par exemple, Bryc et Smoleński [15] ont construit un processus de Markov additif sur un espace d'état dénombrable qui n'admet pas de PGD.

Dans le cas où Σ est fini, nous avons vu que $e^{-n\Lambda(t)} Q(t)^n I$ converge vers un vecteur propre à droite. Mais, dans le cas général, il se peut que $Q(t)^n J$ se comportent différemment suivant que J soit la fonction indicatrice de Σ ou la fonction indicatrice dans un ensemble plus petit A. Ney et Nummelin montrent qu'il faut plutôt utiliser la notion de paramètre de convergence définie par Tweedie [63, 64]. Ils introduisent alors une définition d'ensembles suffisamment petits (s-petits) A pour lesquels $Q(t)^n \mathbf{1}_A$ se comporte comme dans le cas où Σ est fini.

Définition 3.14. *Soit Q un noyau de transition sur un espace polonais Σ. Pour tout distribution initiale ν_0 et toute partie mesurable B de Σ, on définit la série $G^{(\rho)}(\nu_0,\, B)$ par*

$$G^{(\rho)}(\nu_0,\, B) := \sum_{n=0}^{\infty} \rho^n (\nu_0\, Q^n)(B).$$

Si ν_0 et B sont choisis pour qu'il existe un rang m_0 et un réel $0 < \delta \leq 1$ tels que, pour tout $x \in \Sigma$, pour toute partie A mesurable de Σ,

$$\delta \mathbf{1}_B(x) \nu_0(A) \leq Q^{m_0}(x,\, A),$$

alors le rayon de convergence de la série $G^{(\rho)}(\nu_0,\, B)$ est indépendant de ν_0 et de B. Nous appelons ***paramètre de convergence*** *du noyau Q ce rayon de convergence.*

De tels ν_0 et B n'existent pas nécessairement. Ney et Nummelin imposent donc une condition de minoration du noyau pour définir le paramètre de convergence. Ils supposent qu'il existe une distribution μ sur Σ et une famille $(h_a)_{a\in\Sigma}$ de mesures sur $\mathbb{R}^d$ telles que, pour tout $a \in \Sigma$, $A \in \mathcal{B}(\Sigma)$, $B \in \mathcal{B}(\mathbb{R}^d)$,

$$h_a(B)\mu(A) \leq \mathbb{P}_a(X_1 \in A\,;\, f(X_1) \in B).$$

Ils définissent alors $\Lambda(t)$ comme le logarithme du paramètre de convergence du noyau $Q(t)$, où $Q(t)$ est le noyau donné par

$$Q(t)(a,\, B) := \exp(t\, f(a)) Q(a,\, B), \quad a \in \Sigma,\, B \in \mathcal{B}(\Sigma).$$

De plus, quand la chaîne de Markov est une chaîne de Harris, on peut obtenir une équation caractéristique de $\Lambda(t)$.

Chapitre 4

Approximation de la fonction caractéristique

Sommaire

Dans ce chapitre, nous donnons une approximation asymptotique des fonctions caractéristiques d'un processus de Markov additif (S_n) à valeurs dans $\mathbb{Z}^d$. Nous montrons qu'il existe un voisinage de l'origine dans $\mathbb{R}^d$ tel que, pour tout t dans ce voisinage et pour tout état initial $a \in \Sigma$,

$$\mathbb{E}_a\left(\mathrm{e}^{\mathrm{i}\,t\,S_n}\right) \approx G(t)_a \mathrm{e}^{\mathrm{i}\,n\Lambda(\mathrm{i}\,t)}.$$

De plus, l'erreur est exponentiellement négligeable devant cette approximation et les fonctions $t \mapsto G(t)_a$, $t \mapsto \Lambda(\mathrm{i}\,t)$ admettent un développement en série entière en 0. Ce résultat nous permet d'obtenir un développement asymptotique de la loi de $\mathbb{P}_a(S_n = x)$ dans le chapitre 5. De plus, nous montrons, avec une hypothèse supplémentatire (4.1), que la fonction caractéristique de S_n est exponentiellement décroissante avec n, sur $\mathbb{T} = [-\pi\,;\,\pi]^d$ privé d'un voisinage de l'origine.

Le contexte mathématique est identique à celui du chapitre précédent. Nous nous donnons une chaîne de Markov $(X_i)_{i\in\mathbb{N}}$ sur un espace d'états fini Σ et une fonction $f : \Sigma \to \mathbb{R}^d$. Nous étudions le processus de Markov additif $(S_n)_{n\in\mathbb{N}}$ défini par

$$S_n = \sum_{j=0}^{n-1} f(X_j).$$

Nous notons Q le noyau de transition de la chaîne de Markov, que nous supposons irréductible et apériodique. Nous notons π la distribution stationnaire de la chaîne.

Définition 4.1. *Soit J un vecteur de $\mathbb{C}^\Sigma$. Pour tout $z \in \mathbb{C}^d$, nous définissons le noyau $Q(z)$ avec*

$$Q(z)(a,\, b) := Q(a,\, b) \exp z\, f(a), \quad a,\, b \in \Sigma.$$

Pour tout $n \in \mathbb{N}$, $z \in \mathbb{C}^d$, nous définissons

$$J(n,\, z)_a := \mathbb{E}_a(\mathrm{e}^{z\, S_n} J_{X_n}) = (Q(z)^n J)_a.$$

Dans toute la suite, nous fixons $J \in \mathbb{C}^\Sigma$ et $z_0 \in \mathbb{R}^d$. Nous étudions le comportement asymptotique de $J(n, z)$ quand n tend vers $+\infty$, uniformément en z au voisinage du point z_0. Quand z est imaginaire pur et quand J est la fonction indicatrice de Σ, $J(n,\, z)$ est le vecteur des fonctions caractéristiques de S_n pour les différents états initiaux a de la chaîne de Markov.

4.1 Résultats

D'après le chapitre précédent, nous savons que $Q(z_0)$ admet une valeur propre dominante pour tout $z_0 \in \mathbb{R}^d$. Nous voulons étendre ce résultat pour des valeurs complexes de z, afin d'obtenir une approximation de la fonction caractéristique. Malheureusement, les techniques basées sur le théorème de Perron-Frobenius du chapitre 3 ne sont plus utilisables ici : le noyau $Q(z)$ devient complexe quand z est complexe. Fixons $z_0 \in \mathbb{R}^d$. D'après ce qui précède, $Q(z_0)$ admet une valeur propre dominante ρ_0. Alors, nous pouvons utiliser le fait que $Q(z)$ est une perturbation analytique de $Q(z_0)$ et obtenir ainsi une approximation de $\mathbb{E}_a(\exp(z\, S_n))$ lorsque z est au voisinage de z_0. Il vient le résultat suivant.

Proposition 4.2. *Choisissons R parmi les nombres réels positifs, supérieurs au second plus grand module du spectre de $\rho_0^{-1} Q(z_0)$, $R < 1$. Il existe*
- *une constante $c > 0$,*
- *un ouvert V de $\mathbb{C}^d$ contenant z_0,*
- *une fonction holomorphe ρ définie sur V et à valeurs dans $\mathbb{C}$ et*
- *une fonction holomorphe N définie sur V et à valeurs dans les matrices de projections*

tels que $\rho(z)$ est une valeur propre dominante de $Q(z)$ pour tout $z \in V$. Et, pour tout $n \geq 1$, $z \in V$,

$$\left\| \rho(z)^{-n} Q(z)^n - N(z) \right\| \leq c R^n.$$

Dans sa thèse, Mann [40] montre un résultat semblable à cette proposition, mais la convergence de $\rho(z)^{-n} Q(z)^n$ vers $N(z)$ est ponctuelle. Dans notre résultat, cette convergence est uniforme en z. Ceci simplifie beaucoup la démonstration du fait que ρ et N sont holomorphes, par rapport aux estimations de

Mann sur les coefficients du développement en série entière pour montrer que celui-ci converge.

Comme $\rho(0) = 1$, nous pouvons prolonger la définition de Λ de la manière suivante.

Corollaire 4.3. *La fonction $\Lambda = \log \circ \rho$ est bien définie et holomorphe sur un voisinage V de l'origine dans $\mathbb{C}^d$.*

De plus, on obtient une approximation de la fonction caractéristique de S_n au voisinage de l'origine.

Corollaire 4.4. *Il existe une constante c et un voisinage U de l'origine dans $\mathbb{R}^d$ tels que, pour tout $a \in \Sigma$, $n \geq 1$ et $t \in U$,*

$$\left|J(n, \mathrm{i}\, t)_a - \mathrm{e}^{n\Lambda(\mathrm{i}\, t)}\, G(\mathrm{i}\, t)_a\right| \leq c\, R^n\, |\mathrm{e}^{n\Lambda(\mathrm{i}\, t)}\, G(\mathrm{i}\, t)_a|.$$

Le résultat précédent décrit le comportement de la fonction caractéristique au voisinage de l'origine. Nous avons aussi besoin de contrôler ces fonctions caractéristiques quand $n \to +\infty$ en dehors de ce voisinage de 0. Supposons que pour tout vecteur e de la base canonique de $\mathbb{R}^d$, il existe deux états a et b dans Σ, un vecteur x de $\mathbb{Z}^d$ et un rang $n \geq 1$ tels que

$$\begin{cases} \mathbb{P}_a(S_n = x\,;\, X_n = b) > 0 \\ \mathbb{P}_a(S_n = x + e\,;\, X_n = b) > 0. \end{cases} \tag{4.1}$$

Alors, on obtient le résultat suivant.

Proposition 4.5. *Soit $t \in \mathbb{T}$, $t \neq 0$. Si λ est une valeur propre de $Q(\mathrm{i}\, t)$, alors $|\lambda| < 1$.*

Corollaire 4.6. *Pour tout $\epsilon > 0$, il existe A dans $]0\,;\,1[$ et $C > 0$ tels que, pour tout $n \geq 1$ et tout $t \in \mathbb{T}$ vérifiant $\|t\| \geq \epsilon$, on a, pour tout $a \in \Sigma$,*

$$\left|\mathbb{E}_a(\exp(\mathrm{i}\, t S_n))\right| \leq C A^n.$$

4.2 Démonstrations

La démonstration de la proposition 4.2 utilise les lemmes suivants.

Lemme 4.7. *Il existe un voisinage de z_0 dans $\mathbb{C}^d$, des fonctions holomorphes ρ, G et K définies sur V et à valeurs respectivement dans $\mathbb{C}$, $\mathbb{C}^\Sigma$, $\mathbb{C}^\Sigma$ telles que, pour tout $z \in V$, $\rho(z)$ est une valeur propre dominante de $Q(z)$, $K(z)$ et $G(z)$ sont des vecteurs propres respectivement à gauche et à droite de $Q(z)$ pour la valeur propre $\rho(z)$.*

Lemme 4.8. *Soit $F(z)$ l'orthogonal de $K(z)$ dans $\mathbb{C}^\Sigma$. Alors, $F(z)$ est stable par $Q(z)$ et $F(z)$ est un supplémentaire du sous-espace $\mathbb{C}\,G(z)$.*

Lemme 4.9. *Il existe deux applications continues T et M définies sur V et à valeurs dans l'ensemble des matrices carrées de tailles respectivement d et $(d-1)$ telles que, pour tout $z \in V$, $T(z)$ est inversible et*

$$\rho(z)^{-1}Q(z) = T(z) \begin{bmatrix} 1 & 0 \\ 0 & M(z) \end{bmatrix} T(z)^{-1}.$$

Donc, le rayon spectral de $M(z_0)$ est strictement plus petit que 1. Ce qui prouve que nous pouvons choisir R parmi tous les nombres plus grand que le rayon spectral de $M(z_0)$. Ainsi, la proposition 4.2 montre qu'il existe un ouvert $\mathcal{D}$ contenant la droite réelle sur lequel, pour tout $z \in \mathcal{D}$, $Q(z)$ admet une valeur propre dominante.

Démonstration du lemme 4.7. Le polynôme caractéristique de $Q(z)$ est une perturbation analytique du polynôme caractéristique de $Q(z_0)$, qui admet une valeur propre dominante ρ_0 par hypothèse. Donc, au moins sur un voisinage de z_0 dans $\mathbb{C}^d$, une valeur propre est dominante. Notons la $\rho(z)$. De plus ρ est analytique au voisinage de z_0 car ρ_0 est une racine simple.

De même, un vecteur propre à droite (respectivement à gauche) de $Q(z)$ pour la valeur propre $\rho(z)$ est solution d'une perturbation analytique du système linéaire qui donne un vecteur propre à droite (respectivement à gauche) de $Q(z_0)$ pour la valeur propre simple z_0. □

Démonstration du lemme 4.8. Comme $K(z)$ est un vecteur propre à gauche pour $Q(z)$, $F(z)$ est stable par $Q(z)$. Ensuite, il suffit de montrer que $G(z) \notin F(z)$, donc que le produit scalaire $K(z)\,G(z)$ est non nul. Comme $z_0 \in \mathbb{R}^d$, le théorème de Perron-Frobenius montre que les coordonnées de $G(z_0)$ et $K(z_0)$ sont strictement positives. Comme les fonctions $z \mapsto G(z)$ et $z \mapsto K(z)$ sont continues, il existe bien un voisinage de z_0 sur lequel $K(z)\,G(z)$ reste positif. □

Démonstration du lemme 4.9. Choisissons une base $B(z)$ de $F(z)$ qui dépend analytiquement de z sur un voisinage de z_0. Alors $B(z) \cup \{G(z)\}$ dépend analytiquement de z sur un voisinage de z_0. De plus, la matrice de l'endomorphisme $Q(z)$ dans la base $B(z) \cup \{G(z)\}$ est diagonale par blocs, puisque $\mathbb{C}\,G(z)$ et $F(z)$ sont stables par $Q(z)$. □

Démonstration de la proposition 4.2. Soit r_M le rayon spectral de $M(z_0)$. Comme R est plus grand que r_M, pour tout $r \in]r_M\,;\,R[$, le théorème de Householder nous donne une norme sur $\mathbb{C}^\Sigma$ telle que la norme subordonnée de $M(z_0)$ soit

au plus r. Alors, par continuité de M, la norme de $M(z)$ est au plus $R > r$ pour tout z dans un voisinage de z_0. par continuité de T et T^{-1}, les normes de $T(z)$ et $T(z)^{-1}$ sont uniformément bornées pour z dans un voisinage de z_0 par une constante c. Introduisons la matrice

$$N(z) := T(z) \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} T(z)^{-1}.$$

Alors, la norme de $\rho(z)^{-n} Q^n(z) - N(z)$ est au plus $c^2 R^n$. □

Regardons maintenant le comportement de la fonction caractéristique en dehors d'un voisinage de 0.

Démonstration de la proposition 4.5. Le module d'une valeur propre de $Q(\mathrm{i}\,t)$ est au plus 1, puisque, pour tout $a \in \Sigma$,

$$\sum_b |Q(\mathrm{i}\,t)(a,b)| = \sum_b Q(a,b) = 1.$$

Supposons qu'un nombre complexe de module 1 est une valeur propre de $Q(\mathrm{i}\,t)$. Notons $\mathrm{e}^{\mathrm{i}\,s}$ cette valeur propre. Nous voulons montrer que $t = 0$.

Soit Y un vecteur propre à droite de $Q(\mathrm{i}\,t)$ pour la valeur propre $\mathrm{e}^{\mathrm{i}\,s}$. Quitte à multiplier Y par une constante, on peut supposer que 1 est le maximum des $|Y_a|$, $a \in \Sigma$, et choisissons a dans Σ pour que $|Y_a| = 1$. Alors

$$1 = \left|\mathrm{e}^{\mathrm{i}\,s}\, Y_a\right| = \left|\mathrm{e}^{\mathrm{i}\,t\,f(a)} \sum_b Q(a,b)\, Y_b\right| = \left|\sum_b Q(a,b)\, Y_b\right|.$$

Comme la somme sur b des $Q(a,b)$ vaut 1 et comme tout $|Y_b| \leq 1$, il vient que $|Y_b| = 1$ pour tout b dans Σ vérifiant $Q(a,b) \neq 0$. Comme la matrice de transition Q est irréductible, $|Y_a| = 1$ pour tout a dans Σ. On peut donc noter $Y_a =: \mathrm{e}^{\mathrm{i}\,y(a)}$, avec $y(a) \in \mathbb{R}$. Pour tout $n \geq 1$, Y est un vecteur propre à droite de $Q(\mathrm{i}\,t)^n$ pour la valeur propre $\mathrm{e}^{\mathrm{i}\,n\,s}$, c'est-à-dire

$$\mathbb{E}_a\left(\mathrm{e}^{\mathrm{i}\,t\,S_n + \mathrm{i}\,y(X_n)}\right) = \mathrm{e}^{\mathrm{i}\,n\,s + \mathrm{i}\,y(a)}.$$

Ce qui implique que, $\mathbb{P}_a$ presque sûrement,

$$t\,S_n + y(X_n) = n\,s + y(a) \quad \text{modulo } 2\pi.$$

D'après l'hypothèse (4.1), pour tout vecteur e de la base canonique de $\mathbb{Z}^d$, il existe a, b, n and x, tel que les événements $\{S_n = x,\, X_n = b\}$ et $\{S_n = x + e,\, X_n = b\}$ sont tout deux de probabilité non nulle sous $\mathbb{P}_a$. D'où

$$t\,x + y(b) = n\,s + y(a) = t\,(x + e) + y(b) \quad \text{modulo } 2\pi,$$

c'est-à-dire $t\,e = 0$ modulo 2π, pour tout vecteur e de la base canonique. Donc, $t = 0$. □

Démonstration du corollaire 4.6. Soit $r(t)$ le rayon spectral de $Q(\mathrm{i}\,t)$. Comme $Q(\mathrm{i}\,t)$ dépend continûment de t, $r(t)$ dépend continûment de t. La proposition 4.5 implique que $r(t) < 1$ pour tout $t \in \mathbb{T}$ tel que $\|t\| \geq \varepsilon$. Donc le supremum de $r(\cdot)$ sur cet ensemble compact est au plus $A' < 1$. Rappelons que, pour toute matrice M de rayon spectral r_M, la norme de M^n vérifie $(r_M)^{n+o(n)}$ quand $n \to \infty$. Il vient donc que, pour tout $A > A'$, la suite $|\mathbb{E}_a(\mathrm{e}^{\mathrm{i}\,t\,S_n})|/A^n$ est bornée. Ce qui montre la proposition. □

Chapitre 5

Développement de Edgeworth sur $\mathbb{Z}^d$

Sommaire

Nous avons vu, dans le chapitre 2, que la méthode du point selle utilise un développement local limite de la densité de la loi de S_n ou de sa fonction de répartition. Dans ce chapitre, nous démontrons un développement à tous les ordres de la densité de la loi de S_n contre la mesure de comptage de $\mathbb{Z}^d$, c'est-à-dire de $\mathbb{P}_a(S_n = x)$, quand n tend vers $+\infty$. La différence entre $\mathbb{P}_a(S_n = x)$ et le développement asymptotique que nous obtenons est bornée uniformément en x. Ce développement est construit à partir de la densité de la loi normale qui apparaît dans le théorème central limite des processus de Markov additif.

Nous nous plaçons dans le contexte mathématique suivant. Nous nous donnons une chaîne de Markov $(X_i)_{i\in\mathbb{N}}$ sur un espace d'états Σ et une fonction $f : \Sigma \to \mathbb{Z}^d$. Nous étudions le processus de Markov additif $(S_n)_{n\in\mathbb{N}}$ sur $\mathbb{Z}^d$ défini par

$$S_n = \sum_{j=0}^{n-1} f(X_j).$$

Nous notons Q le noyau de transition de la chaîne de Markov, que nous supposons irréductible et apériodique. Nous supposons que Σ est fini et nous notons π la distribution stationnaire de la chaîne.

5.1 Définitions préliminaires

Notons $m := \pi f$ la dérive de S_n et Γ sa matrice de covariance asymptotique, c'est-à-dire

$$\Gamma := n^{-1}\mathrm{Var}(S_n).$$

Rappelons que $\Lambda(z)$, $z \in \mathbb{C}^d$, est, par définition, le logarithme de la valeur propre dominante de $Q(z)$, et que Λ est une fonction holomorphe en 0. De plus, il existe un vecteur propre à droite $G(z)$ pour la valeur propre $\mathrm{e}^{\Lambda(z)}$ de $Q(z)$. Et G est également une fonction holomorphe à l'origine.

Définition 5.1. *Les développements de Taylor de Λ et G à l'origine s'écrivent*

$$\Lambda(z) = \sum_{j=1}^{+\infty} \Lambda^{(j)}(z), \quad G(z)_a = \sum_{j=0}^{+\infty} G(z)_a^{(j)},$$

pour tout $a \in \Sigma$, où $\Lambda^{(j)}$ et $G(\cdot)_a^{(j)}$ sont des polynômes homogènes de degrés j. Ainsi, pour tout $a \in \Sigma$,

$$G(0)_a = \pi\, J, \quad \Lambda^{(1)}(z) = m\, z, \quad \Lambda^{(2)}(z) = z\, \Gamma\, z/2.$$

Enfin, nous notons $L := \Lambda - \Lambda^{(1)} - \Lambda^{(2)}$.

Proposition 5.2. *L'égalité*

$$P(u,\, z) := \exp\left(\frac{L(uz)}{u^2}\right) G(uz), \quad u \in \mathbb{C},\, z \in \mathbb{C}^d$$

définie une fonction vectorielle $P(\cdot,\, z)$, analytique à l'origine. Son développement en puissance de u s'écrit

$$P(u,\, z) =: \sum_{k=0}^{\infty} P_k(z)\, u^k.$$

Chaque fonction $z \mapsto P_k(z)_a$ est une fonction polynomiale en $z = (z_1, \ldots,\, z_d)$, de degré au plus $(3k)$ qui s'écrit à l'aide d'un nombre fini de $\Lambda^{(j)}(z)$ et $G(z)_a^{(j)}$.

Exemple 5.1. Les premiers P_k sont donnés par $P_0(z) = G(z)^{(0)} = G(0)$, donc $P_0(z)_a = \pi\, J$ pour tout $a \in \Sigma$, et

$$P_1(z)_a = \Lambda^{(3)}(z) G(z)_a^{(0)} + G(z)_a^{(1)}.$$

Définition 5.3. *Soit $k \in \mathbb{N}$ et $a \in \Sigma$. Notons φ_Γ la densité de la loi gaussienne centrée de matrice de covariance Γ. Nous définissons la fonction ψ_a^k sur $\mathbb{R}^d$ par*

$$\psi_a^k := P_k(D)_a \varphi_\Gamma.$$

5.2 Résultats

Nous obtenons un développement de Edgeworth pour la densité de la loi de S_n contre la mesure de comptage de $\mathbb{Z}^d$ à tous les ordres. Ce théorème répond à la question posée dans la remarque 3.4.

Théorème 5.4. *Il existe une suite de constante finie $(C_k)_{k\in\mathbb{N}}$, qui dépendent de J telles que, pour tout $k \in \mathbb{N}$, $n \in \mathbb{N}$, $a \in \Sigma$ et $x \in \mathbb{Z}^d$, on ait*

$$\left| \mathbb{E}_a \left[J_{X_n} \,;\, S_n = x \right] - n^{-d/2} \sum_{j=0}^{k} n^{-j/2}\, \psi_a^j \left(\frac{x - n\,m}{\sqrt{n}} \right) \right| \leq C_k\, n^{-(d+k+1)/2}.$$

Remarque 5.1. Lorsque x est loin de $n\,m$, il se peut que

$$\mathbb{E}_a \left[J_{X_n} \,;\, S_n = x \right] \quad \text{et son développement} \quad n^{-d/2} \sum_{j=0}^{k} n^{-j/2}\, \psi_a^j \left(\frac{x - n\,m}{\sqrt{n}} \right)$$

soient chacun négligeable devant $n^{-(d+k+1)/2}$. Prenons un exemple simple, en choisissant $(X_j)_{j\in\mathbb{N}}$ i.i.d. de loi de Bernoulli $(1-p)\mathbf{1}_0 + p\,\mathbf{1}_1$, $x = 0$ et $J_a = 1$ pour tout a. Alors $\mathbb{P}(S_n = 0) = (1-p)^n$ décroit exponentiellement. Pour $k = 0$, ψ^0 est la densité de la loi gaussienne centrée de variance $p(1-p)$ et c'est la seule fonction ψ_a^j qui intervient dans le développement. Donc $n^{-1/2}\psi^0(-\sqrt{n}p)$ est égal, à une constante multiplicative près, à $n^{-1/2}\exp(-np/2(1-p))$, qui décroit aussi exponentiellement avec n. Il existe donc bien C tel que la différence des deux soit inférieur à Cn^{-1} pour toute valeur de n comme l'affirme le théorème 5.4. Mais, $n^{-1/2}\psi^0(-\sqrt{n}p)$ n'est pas équivalent à $\mathbb{P}(S_n = 0)$. En revanche, pour x voisin de $n\,m$, $\psi_a^0(n^{-1/2}(x - n\,m))$ est voisin de $\psi_a^0(0)$, constante strictement positive, et $C\,n^{-(d+1)/2}$ est bien négligeable devant le premier terme du développement.

Remarque 5.2. Les constantes C_k ne dépendent pas de x, c'est-à-dire l'inégalité obtenue dans le théorème ci-dessus est uniforme en x.

5.3 Démonstration du développement

5.3.1 Notations

Écrivons les deux termes du membres de gauche du théorème 5.4 comme des transformées de Fourier. Notons $\mathbb{T} := [-\pi\,;\,\pi]^d$. D'un côté, nous avons

$$\mathbb{E}_a \left[J_{X_n} \,;\, S_n = x \right] = (2\pi)^{-d} \int_{\mathbb{T}} J(n, \mathrm{i}\,t)_a\, \mathrm{e}^{-\mathrm{i}\,t\,x}\, \mathrm{d}t.$$

De l'autre côté, pour tout $j \in \mathbb{N}$, on a par la formule d'inversion de Fourier,

$$\begin{aligned} n^{-d/2}\,\psi_a^j(y) &= (2\pi)^{-d} \int_{\mathbb{R}^d} P_j(\mathrm{i}\,t)_a\,\mathrm{e}^{-t\,\Gamma\,t/2-\mathrm{i}\,t\,y}\,\frac{\mathrm{d}t}{n^{d/2}} \\ &= (2\pi)^{-d} \int_{\mathbb{R}^d} P_j(\mathrm{i}\,t\,\sqrt{n})_a \mathrm{e}^{-n\,t\,\Gamma\,t/2-\mathrm{i}\,t\,y\,\sqrt{n}}\,\mathrm{d}t. \end{aligned}$$

Rappelons que les P_j *ne* sont *pas* des polynômes homogènes de degré j. En utilisant ceci pour $y := (x - n\,m)/\sqrt{n}$, on voit que, à un facteur $(2\pi)^d$ près, le membre de gauche du théorème 5.4 est majoré par la somme des trois intégrales $I_1(n)$, $I_2(n)$ et $I_3(n)$, définies par

$$\begin{aligned} I_1(n) &:= \int_{\|t\|\le\varepsilon} \left| J(n,\mathrm{i}\,t)_a - \sum_{j=0}^{k} n^{-j/2}\,P_j(\mathrm{i}\,t\,\sqrt{n})_a\,\mathrm{e}^{-n\,t\,\Gamma\,t/2+\mathrm{i}\,n\,t\,m} \right| \mathrm{d}t, \\ I_2(n) &:= \int_{\|t\|\ge\varepsilon,\,t\in\mathbb{T}} |J(n,\mathrm{i}\,t)_a|\;\mathrm{d}t, \\ I_3(n) &:= \int_{\|t\|\ge\varepsilon} \left| \sum_{j=0}^{k} n^{-j/2}\,P_j(\mathrm{i}\,t\,\sqrt{n})_a\,\mathrm{e}^{-n\,t\,\Gamma\,t/2} \right| \mathrm{d}t, \end{aligned}$$

où nous avons annulé les facteurs inutiles $\mathrm{e}^{-\mathrm{i}\,t\,x}$ dans $I_1(n)$ et $I_2(n)$, et $\mathrm{e}^{-\mathrm{i}\,t\,y\sqrt{n}}$ dans $I_3(n)$. Dans les prochaines étapes de la démonstrations, nous montrons d'abord que $I_2(n)$ et $I_3(n)$ sont exponentiellement petits, puis nous utilisons l'approximation de $J(n,\,\mathrm{i}\,t)$ du corollaire 4.6 pour remplacer $J(n,\,\mathrm{i}\,t)$ par $G(\mathrm{i}\,t)\mathrm{e}^{n\Lambda(\mathrm{i}\,t)}$. Et finalement, nous montrons que $I_1(n)$ est borné par une puissance de n.

5.3.2 Majoration de $I_2(n)$ et $I_3(n)$

D'après le corollaire 4.6 de la proposition 4.5, $|J(n,\mathrm{i}\,t)_a| \le C\,A^n$ avec $A < 1$. Donc, $I_2(n)$ est exponentiellement petit, puisque

$$I_2(n) \le C\,A^n \mathrm{Vol}(\mathbb{T}).$$

Concernant $I_3(n)$, P_j est un polynôme de degré au plus $(3j)$, donc, pour tout $\|z\| \ge \epsilon$, $|P_j(z)_a| \le c_{j,a}\,\|z\|^{3j}$. En appliquant ceci avec $z := \mathrm{i}\,t\,\sqrt{n}$, il vient

$$I_3(n) \le \sum_{j=0}^{k} c_{j,a} n^j \int_{\|t\|\ge\varepsilon} \mathrm{e}^{-n\,t\,\Gamma\,t/2}\,\|t\|^{3j}\,\mathrm{d}t.$$

Comme Γ est définie positive, $t\,\Gamma\,t \ge g\,\|t\|^2$, où $g > 0$ est la plus petite valeur propre de Γ. D'où, pour tout $n \ge 1$,

$$\int_{\|t\|\ge\varepsilon} \mathrm{e}^{-n\,t\,\Gamma\,t/2}\,\|t\|^{3j}\,\mathrm{d}t \le \mathrm{e}^{-n\,g\,\varepsilon^2/4} \int \mathrm{e}^{-t\,\Gamma\,t/4}\,\|t\|^{3j}\,\mathrm{d}t.$$

La dernière intégrale ci-dessus est convergente, donc $I_3(n)$ est exponentiellement petit.

5.3.3 Approximation de $J(n, \mathrm{i}\,t)$ dans $I_1(n)$

D'après le corollaire 4.4, si ε est suffisamment petit, il vient, pour tout $\|t\| \leq \varepsilon$,

$$\left|J(n, \mathrm{i}\,t)_a - G(\mathrm{i}\,t)_a\,\mathrm{e}^{n\,\Lambda(\mathrm{i}\,t)}\right| \leq c\,R^n\left|\mathrm{e}^{n\,\Lambda(\mathrm{i}\,t)}\right|.$$

D'où $I_1(n) \leq I_4(n) + c\,R^n\,I_5(n)$, avec

$$I_4(n) := \int_{\|t\|\leq\varepsilon}\left|G(\mathrm{i}\,t)_a\,\mathrm{e}^{n\,\Lambda(\mathrm{i}\,t)} - \sum_{j=0}^{k} n^{-j/2}\,P_j(\mathrm{i}\,t\,\sqrt{n})_a\,\mathrm{e}^{-n\,t\,\Gamma\,t/2+\mathrm{i}\,t\,m\,n}\right|\,\mathrm{d}t$$

$$I_5(n) := \int_{\|t\|\leq\varepsilon}\left|\mathrm{e}^{n\,\Lambda(\mathrm{i}\,t)}\right|\,\mathrm{d}t.$$

Comme $\Lambda^{(0)}(z) = 0$, $\Lambda^{(1)}(z) = m\,z$ et $\Lambda^{(2)}(z) = z\,\Gamma\,z/2$,

$$\mathrm{e}^{n\,\Lambda(\mathrm{i}\,t)} = \mathrm{e}^{-n\,t\,\Gamma\,t/2+\mathrm{i}\,t\,m\,n+n\,L(\mathrm{i}\,t)}, \quad L := \sum_{j=3}^{\infty}\Lambda^{(j)}.$$

Nous pouvons annuler les termes $\mathrm{e}^{\mathrm{i}\,t\,m\,n}$ dans $I_4(n)$ et $I_5(n)$, et factoriser les termes $\mathrm{e}^{-n\,t\,\Gamma\,t/2}$ dans $I_4(n)$. Nous obtenons

$$I_4(n) = \int_{\|t\|\leq\varepsilon}\left|G(\mathrm{i}\,t)_a\,\mathrm{e}^{n\,L(\mathrm{i}\,t)} - \sum_{j=0}^{k} n^{-j/2}\,P_j(\mathrm{i}\,t\,\sqrt{n})_a\right|\,\mathrm{e}^{-n\,t\,\Gamma\,t/2}\,\mathrm{d}t$$

$$I_5(n) = \int_{\|t\|\leq\varepsilon}\left|\mathrm{e}^{n\,L(\mathrm{i}\,t)}\right|\,\mathrm{e}^{-n\,t\,\Gamma\,t/2}\,\mathrm{d}t.$$

5.3.4 Majoration de $I_1(n)$

Nous utilisons le lemme élémentaire d'analyse complexe suivant (voir, par exemple, [16]).

Lemme 5.5. *Supposons que la fonction F est analytique à l'origine dans $\mathbb{C}^d$, que le développement de Taylor de F autour de l'origine s'écrive*

$$F(z) = \sum_{j=0}^{+\infty} F^{(j)}(z),$$

où $F^{(j)}$ est un polynôme homogène de degré j, et supposons que ce développement converge absolument pour $\|z\| \leq 1/\alpha$, où α est une constante strictement positive. Alors, (i) il existe un réel positif β tel que, pour tout $j \in \mathbb{N}$ et tout $z \in \mathbb{C}^d$ vérifiant $\|z\| \leq 1/\alpha$,

$$\left|F^{(j)}(z)\right| \leq \beta\,(\alpha\,\|z\|)^j,$$

et, (ii) pour tout z vérifiant $\|z\| \le 1/(2\alpha)$ et tout $\ell \in \mathbb{N}$,

$$\left| F(z) - \sum_{j=0}^{\ell} F^{(j)}(z) \right| \le 2\beta(\alpha\|z\|)^{\ell+1}.$$

En appliquant la partie (ii) du lemme 5.5 à $F := \Lambda$ and $\ell := 2$, si ε est suffisamment petit, alors $|L(\mathrm{i}\,t)| \le c\,\|t\|^3$ pour tout t tel que $\|t\| \le \varepsilon$, où c est une constante finie. Comme Γ est définie positive, $t\,\Gamma\,t \ge g\,\|t\|^2$ avec une constante g strictement positive. Donc, pour $\varepsilon \le g/(4c)$ et $\|t\| \le \varepsilon$, $|L(\mathrm{i}\,t)| \le t\,\Gamma\,t/4$. Ce qui donne

$$I_5(n) \le \int_{\|t\|\le\varepsilon} \mathrm{e}^{-n\,g\,\|t\|^2/4}\,\mathrm{d}t = O(n^{-d/2}).$$

Quant à $I_4(n)$, la proposition 5.2 avec $u = n^{-1/2}$ et $z = \mathrm{i}\,t\sqrt{n}$ donne

$$G(\mathrm{i}\,t)_a\,\mathrm{e}^{n\,L(\mathrm{i}\,t)} = \sum_{j\ge 0} n^{-j/2}\,P_j(\mathrm{i}\,t\,\sqrt{n})_a.$$

Donc,

$$I_4(n) = \int_{\|t\|\le\varepsilon} \left| \sum_{j\ge k+1} n^{-j/2}\,P_j(\mathrm{i}\,t\,\sqrt{n})_a \right| \mathrm{e}^{-n\,t\,\Gamma\,t/2}\,\mathrm{d}t.$$

Nous utilisons le lemme 5.6 ci-dessous pour majorer chaque P_j. La preuve de ce lemme est dans le paragraphe 5.3.5.

Lemme 5.6. *Il existe deux constantes strictement positives α et β telles que, pour tout $j \in \mathbb{N}$ et tout z vérifiant $\|z\| \le 1/(2\alpha)$, on ait*

$$|P_j(z)_a| \le \beta\,(2y)^j \sum_{\ell=0}^{j} (\beta\,y^2)^\ell/\ell!, \qquad y := \alpha\|z\|.$$

Nous utilisons le fait que $t\,\Gamma\,t \ge g\,\|t\|^2$ et le changement de variables $s := \alpha\|t\|\sqrt{n}$ de la manière suivante :

$$\int_{\|t\|\le\epsilon} g(\|t\|)\,\mathrm{d}t = \alpha^{-d} n^{-d/2} \rho_d \int_{s=0}^{\alpha\epsilon\sqrt{n}} s^{d-1}\,g(s/\alpha\sqrt{n})\,\mathrm{d}s,$$

où ρ_d est l'aire de la sphère unité de $\mathbb{R}^d$, quelle que soit la fonction g à valeurs réelles positives. En regardant séparément les indices ℓ du majorant de P_j donné par le lemme 5.6 tels que $\ell \le k$ et les indices ℓ tels que $\ell \ge k+1$, il vient

$$I_4(n) \le \rho_d\,(2/\alpha)^d\,\beta\,n^{-d/2}(I_6(n) + I_7(n)),$$

où

$$I_6(n) := \sum_{\ell=0}^{k} I_8(\ell, n), \quad I_7(n) := \sum_{\ell=k+1}^{\infty} I_8(\ell, n),$$

et

$$I_8(\ell, n) := \sum_{j \geq \max\{k+1,\, \ell\}} \int_0^{\alpha\varepsilon\sqrt{n}} (2s/\sqrt{n})^j \, ((\beta\, s^2)^\ell/\ell!)\, s^{d-1}\, \mathrm{e}^{-gs^2/(2\alpha^2)}\, \mathrm{d}s,$$

À partir de maintenant, nous supposons que $\alpha\varepsilon \leq 1/4$. Alors $2s/\sqrt{n} \leq 1/2$ uniformément sur l'intervalle d'intégration $[0\,;\, \alpha\varepsilon\sqrt{n}]$. D'où, pour tout i,

$$\sum_{j=i}^{\infty} (2s/\sqrt{n})^j \leq 2(2s/\sqrt{n})^i. \tag{5.1}$$

En utilisant (5.1) avec $i := k+1$ pour majorer les termes $I_8(\ell, n)$ tels que $\ell \leq k$, il vient

$$I_6(n) \leq 2^{k+2} n^{-(k+1)/2} \sum_{\ell=0}^{k} \int_0^{+\infty} ((\beta s^2)^\ell/\ell!) s^{k+d}\, \mathrm{e}^{-gs^2/(2\alpha^2)}\, \mathrm{d}s.$$

Il n'y a qu'un nombre fini d'intégrales dans cette dernière somme, donc il existe un réel $c_6(k) < +\infty$, qui ne dépend pas de n, tel que

$$I_6(n) \leq c_6(k)\, n^{-(k+1)/2}.$$

Quant à $I_7(n)$, nous utilisons (5.1) avec $i := \ell$ pour majorer les termes $I_8(\ell, n)$ tels que $\ell \geq k+1$. De plus,

$$\sum_{\ell=k+1}^{\infty} 2(2s/\sqrt{n})^\ell ((\beta\, s^2)^\ell/\ell!) \leq 2(2s/\sqrt{n})^{k+1} (\beta\, s^2)^{k+1}\, \mathrm{e}^{(2s/\sqrt{n})\beta\, s^2}.$$

Comme la dernière exponentielle vaut au plus $\mathrm{e}^{2\alpha\varepsilon\beta\, s^2}$,

$$I_7(n) \leq 2(2\beta/\sqrt{n})^{k+1} \int_0^{+\infty} s^{2k+d+1}\, \mathrm{e}^{-\gamma\, s^2/2}\, \mathrm{d}s.$$

où $\gamma := g/\alpha^2 - 4\alpha\varepsilon\beta$. Si ε est suffisamment petit, γ est strictement positif et les intégrales ci-dessus sont convergeantes. D'où, $I_7(n) \leq c_7(k)\, n^{-(k+1)/2}$.

Bref, $I_4(n)$, donc $I_1(n)$, donc le membre de gauche du théorème 5.4 sont majorés par des multiples de $n^{-(k+d+1)/2}$.

5.3.5 Démonstration du lemme 5.6

D'après la partie (i) du lemme 5.5, il existe deux constantes positives α et β telles que, pour tout $i \in \mathbb{N}$,

$$|G(z)_a^{(i)}| \leq \beta\, y^i, \quad |\Lambda^{(i)}(z)| \leq \beta\, y^i, \quad y := \alpha\|z\|.$$

Dans la proposition 5.2, le développement de l'exponentielle pour $z \in \mathbb{C}^d$ fixé, donne

$$\mathrm{e}^{L(uz)/u^2} = \sum_{\ell=0}^{\infty} (L(uz)/u^2)^\ell\, u^\ell/\ell!.$$

De plus,

$$(L(uz)/u^2)^\ell = \sum_{i_1,\ldots,i_\ell \geq 0} u^{i_1+\cdots+i_\ell}\, \lambda(i_1,\ldots,i_\ell)(z),$$

où

$$\lambda(i_1,\ldots,i_\ell)(z) := \Lambda^{(i_1+3)}(z)\cdots\Lambda^{(i_\ell+3)}(z).$$

En extrayant le terme en u^j du produit dans la proposition 5.2, il vient

$$P_j(z)_a = \sum G(z)_a^{(i)}\, \lambda(i_1,\ldots,i_\ell)(z)/\ell!,$$

où l'on somme sur tout les entiers i, ℓ, i_1, …, i_ℓ, vérifiant

$$i + \ell + i_1 + \cdots + i_\ell = j.$$

Comme $|\Lambda^{(i)}(z)| \leq \beta\, y^i$,

$$|\lambda(i_1,\ldots,i_\ell)(z)| \leq \beta^\ell\, y^{i_1+\cdots+i_\ell+3\ell} = \beta^\ell\, y^{j-i+2\ell}.$$

D'où,

$$|P_j(z)_a| \leq \sum_{\ell \leq j} \beta^{\ell+1}\, y^{j+2\ell}\, n_j(\ell)/\ell!,$$

où $n_j(\ell)$ est le nombre de ℓ-uples $(i_1,\ldots,i_\ell)$ pour lesquels il existe un entier i vérifiant $i + \ell + i_1 + \cdots + i_\ell = j$, c'est-à-dire, pour les ℓ-uples tels que

$$\ell + i_1 + \cdots + i_\ell \leq j.$$

Le lemme 5.7 ci-dessous montre que $n_j(\ell) = \binom{j}{\ell}$. Avec la majoration $n_j(\ell) \leq 2^j$, le résultat du lemme 5.6 est démontré.

Lemme 5.7. $n_j(\ell) = \binom{j}{\ell}$.

Démonstration. Supposons que $n_j(\ell)$ est le cardinal de l'ensemble des ℓ-uples $N_j(\ell)$. Considérons l'application qui associe à chaque ℓ-uple $(i_1,\ldots,i_\ell)$ tel que $i_1 = 0$ le $(\ell-1)$-uple $(i_2,\ldots,i_\ell)$ et à chaque ℓ-uple tel que $i_1 \geq 1$ le ℓ-uple $(i_1-1,\ldots,i_\ell)$. Les images des ℓ-uples de $N_j(\ell)$ du premier type parcourent $N_{j-1}(\ell-1)$, et les images des ℓ-uples du second type parcourent $N_{j-1}(\ell)$. Comme l'application est injective, nous avons montré que $n_j(\ell) = n_{j-1}(\ell-1) + n_{j-1}(\ell)$. De plus $n_0(\ell) = 1$ pour tout $\ell \geq 1$, d'où le résultat. □

Chapitre 6

Estimations exactes de grandes déviations

Sommaire

Soit (S_n) un processus de Markov additif sur $\mathbb{Z}^d$. Nous voulons étudier $\mathbb{P}_a(S_n \in nB)$ où B est un produit de demi-droites. Pour que $\{S_n \in nB\}$ soit un événement exceptionnel, on suppose, quitte à changer le signe des coordonnées de f, que $B = \{x \in \mathbb{R}^d\,;\, x \geq v\}$ avec $v > m$, m étant la dérive de S_n. Le principe de grandes déviations pour les chaînes de Markov montre que le comportement à l'échelle logarithmique de $\mathbb{P}_a(S_n \in nB)$ est donné par la valeur de la fonction de taux en v, qui est le point dominant B. C'est donc le comportement local de S_n autour de nv qui dicte la valeur limite de $n^{-1}\log \mathbb{P}(S_n \in nB)$. Pour avoir plus de précision sur le comportement asymptotique de cette probabilité, en particulier pour avoir un équivalent simple, nous avons besoin de connaître précisément la position de nB par rapport au réseau $\mathbb{Z}^d$ sur lequel S_n prend ses valeurs. Nous introduisons donc la suite $(s_n)_{n\in\mathbb{N}}$, à valeurs dans $\mathbb{Z}^d$ qui vérifie $nB \cap \mathbb{Z}^d = \{x \in \mathbb{Z}^d\,;\, x \geq s_n\}$, i.e. $s_n = \lceil nv \rceil$, voir figure 6.1. En fait, nous obtenons un développement asymptotique de $\mathbb{P}_a(S_n \geq s_n)$ et $\mathbb{P}_a(S_n = s_n)$ pour toutes les suites $(s_n)_{n\in\mathbb{N}}$ de $\mathbb{Z}^d$ qui vérifient $s_n = nv + o(\sqrt{n})$.

Nous rassemblons les résultats des chapitres précédents pour obtenir des développements asymptotiques de grandes déviations de (S_n). Cette méthode ressemble dans les grandes lignes à la méthode du point selle que nous avons présentée dans le chapitre 2. Nous obtenons deux développements asymptotiques, le premier pour $\mathbb{P}_a(S_n = s_n)$ dans le théorème 6.1 et le second pour $\mathbb{P}_a(S_n \geq s_n)$ dans le théorème 6.4 où (s_n) est une suite de $\mathbb{Z}^d$ fixée, $s_n = nv + o(n^{1/2})$. En particulier, nous en déduisons des équivalents, qui sont énoncés dans les corollaires 6.2 et 6.5.

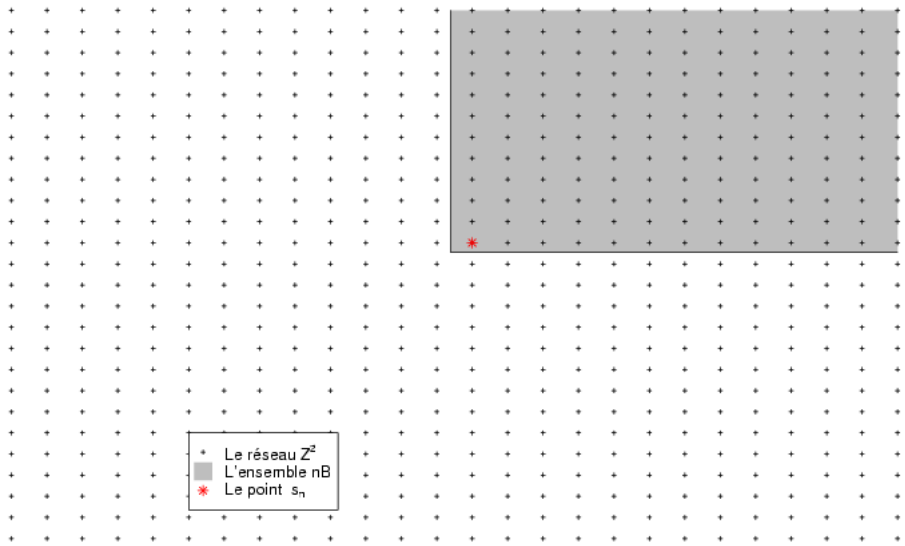

FIG. 6.1 – Exemple de position de l'ensemble $nB \subset \mathbb{R}^2$ par rapport au réseau $\mathbb{Z}^2$, où S_n prend ses valeurs, quand $B = \prod_j [v_j\,;\,+\infty[$. Le point $s_n = \lceil nv \rceil$ de $\mathbb{Z}^2$ spécifie la position de nB par rapport au réseau sur lequel S_n prend ses valeurs.

6.1 Premier type de développements

Soit $X = (X_j)_{j\in\mathbb{N}}$ une chaîne de Markov sur un espace d'état fini Σ et $f : \Sigma \to \mathbb{Z}^d$. Nous nous intéressons aux grandes déviations de

$$S_n := \sum_{j=0}^{n-1} f(X_j).$$

Nous notons Q la matrice de transition de la chaîne de Markov, que nous supposons irréductible et apériodique. Elle admet donc une unique distribution stationnaire π. De plus, nous supposons que, pour tout vecteur e de la base canonique de $\mathbb{Z}^d$, il existe $n \in \mathbb{N}$, un vecteur x de $\mathbb{Z}^d$, et deux états a, $b \in \Sigma$ tels que

$$\begin{cases} \mathbb{P}_a(S_n = x\,;\, X_n = b) > 0 & \text{et} \\ \mathbb{P}_a(S_n = x + e\,;\, X_n = b) > 0 & . \end{cases} \tag{6.1}$$

Nous notons $m := \lim n^{-1}\mathbb{E}_a(S_n) = \pi f$ la dérive de S_n. Nous avons montré, dans le chapitre 3 que pour tout $t \in \mathbb{R}^d$,

$$\Lambda(t) := \lim n^{-1} \log \mathbb{E}_a(\exp t\, S_n)$$

est la valeur propre dominante du noyau $Q(t)$, défini par

$$Q(t)(a,\, b) := Q(a,\, b)\exp(t\, f(a)).$$

De plus, il existe un vecteur propre à droite $G(t)$ de coordonnées strictement positives, qui permet de définir le noyau de transition, dit noyau twisté,

$$Q^{(t)}(a,\, b) := Q(t)(a,\, b)G(t)_b G(t)_a^{-1}.$$

Nous notons $\mathbb{P}^{(t)}$ la mesure de probabilité qui fait de X une chaîne de Markov de noyau de transition $Q^{(t)}$. Ce noyau hérite du caractère irréductible et apériodique de Q. On note $\pi(t)$ sa distribution stationnaire. Nous choisissons $G(t)$ pour que $\pi(t)(G(t)^{-1}) = 1$ (voir l'équation (3.1) dans le chapitre 3). De plus, nous rappelons que Λ' définit un homéomorphisme de $\mathbb{R}^d$ sur l'intérieur de $\mathcal{E}$, l'enveloppe convexe de $f(\Sigma)$, et nous notons Ξ l'homéomorphisme réciproque. Sous $\mathbb{P}^{(t)}$, la dérive de S_n est $\Lambda'(t) = \pi(t)\, f$ et la matrice de covariance vérifie

$$\text{var } S_n \sim n\Gamma(t), \quad \text{où } \Gamma(t) := \text{Hess}\,\Lambda(t)$$

quand $n \to \infty$.

Fixons maintenant une suite (s_n) de $\mathbb{Z}^d$ qui vérifie

$$s_n = nv + o(n^{1/2})$$

pour un certain v dans l'intérieur de $\mathcal{E}$ et fixons $k \in \mathbb{N}$. Nous voulons étudier $\mathbb{P}_a(S_n = s_n)$. Quitte à changer le signe des coordonnées de f, on peut supposer que $v > m$. Choisissons $t = \Xi(v)$ de telle sorte que la dérive de S_n sous $\mathbb{P}^{(t)}$ soit égale à v. Alors, la proposition 3.12 qui relie la loi de S_n sous $\mathbb{P}^{(t)}$ à la loi de S_n sous $\mathbb{P}$ montre que

$$\mathbb{P}_a(S_n = s_n) = \mathrm{e}^{-n\,\Lambda^\star(v) - t\,(s_n - nv)}\, G(t)_a\, D_n$$

avec

$$D_n := \mathbb{E}_a^{(t)}[J_{X_n}\,;\, S_n = s_n], \quad J_b := 1/G(t)_b.$$

Nous pouvons alors utiliser le théorème 5.4 du chapitre 5 sur la chaîne twistée pour $x := s_n$. Pour énoncer ce théorème, nous avons introduit les fonctions ψ_a^j dans la définition 5.3, comme des combinaisons linéaires de dérivées partielles de la densité de la loi gaussienne centrée de matrice de covariance $\Gamma(t)$, les coefficients de ces combinaisons linéaires étant donnés par les développements en séries entières de Λ et G à l'origine. De plus, l'hypothèse (6.1) passe à la chaîne de Markov twistée. Nous sommes donc dans les hypothèses du théorème 5.4 et nous obtenons

$$\left| D_n - n^{-d/2} \sum_{j=0}^{k} n^{-j/2} u_n^j \right| \le C_k n^{-(k+d+1)/2}, \qquad \text{où } u_n^j := \psi_a^j\left(\frac{s_n - nv}{\sqrt{n}}\right).$$

Rappelons que les fonctions ψ_a^j sont bornées sur $\mathbb{R}^d$, donc pour tout j, la suite $(u_n^j)_{n\in\mathbb{N}}$ est bornée. Ce qui donne le développement asymptotique suivant.

Théorème 6.1. *Soit $k \in \mathbb{N}$. Quand $n \to +\infty$,*

$$\mathbb{P}_a(S_n = s_n) = n^{-d/2} \mathrm{e}^{-t\,(s_n - nv) - n\Lambda^\star(v)} \left(\sum_{j=0}^{k} n^{-j/2} u_n^j + O(n^{-(k+1)/2}) \right).$$

Avec $k = 0$, on obtient les deux corollaires suivants. On note $\langle \Delta \rangle$ la valeur en 0 de la densité de loi gaussienne centrée de matrice de covariance Δ.

Corollaire 6.2. *Quand $n \to +\infty$, si $s_n = nv + o(n^{1/2})$,*

$$\mathbb{P}_a(S_n = s_n) \sim C(a, v)\, n^{-d/2}\, \mathrm{e}^{-t\,(s_n - nv) - n\Lambda^\star(v)},$$

où $t = \Xi(v)$ est défini par

$$\Lambda^\star(v) = t\, v - \Lambda(t)$$

et $C(a, v)$ est la constante finie donnée par

$$C(a, v) := \langle \Gamma(t) \rangle\, G(t)_a.$$

Démonstration. En effet, avec $k = 0$, la seule fonction ψ_a^j qui apparaît dans le développement est $\psi_a^0(\cdot) = (\pi(t)\, J)\, \varphi_{\Gamma(t)}(\cdot)$, et l'espérance asymptotique sous $\mathbb{P}_a^{(t)}$ est $(\pi(t)\, f) = v$. D'où

$$n^{d/2}\, D_n = (\pi(t)\, J)\, \varphi_{\Gamma(t)}(s'_n) + O(1/\sqrt{n}), \quad s'_n := (s_n - nv)/\sqrt{n}.$$

De plus,

$$\varphi_{\Gamma(t)}(s'_n) = \langle \Gamma(t) \rangle + O(s'_n)$$

et $\pi(t)(G(t)^{-1}) = 1$. Comme $s'_n = o(1)$, le corollaire est bien démontré. □

De même, on obtient le résultat suivant lorsqu'on impose un contrôle plus précis sur s_n.

Corollaire 6.3. *Supposons que $s_n = nv + O(1)$. Alors, quand $n \to +\infty$,*

$$\mathbb{P}_a(S_n = s_n) = \big(C(a, v) + O(n^{-1/2}) \big)\; n^{-d/2}\, \mathrm{e}^{-t\,(s_n - nv) - n\Lambda^\star(v)}.$$

Remarque 6.1. Comme s_n appartient à $\mathbb{Z}^d$, $s_n = nv + O(1)$ est le contrôle le plus fort que l'on puisse imposer sur s_n en toute généralité. En particulier, les suites $s_n = \lfloor nv \rfloor$ et $s_n = \lceil nv \rceil$ vérifient cette condition.

Remarque 6.2. Nous avons constaté que le développement donné dans le théorème 5.4 est parfois négligeable par rapport à l'erreur $O(n^{-(k+d+1)/2})$, suivant la position de x en fonction de la valeur moyenne de S_n. Une telle situation n'arrive jamais avec le théorème 6.1 énoncé ci-dessus. En effet, le corollaire 6.2 montre que u_n^0 tend vers une constante strictement positive. Donc

$$\sum_{j=0}^{k} n^{-j/2} u_n^j$$

n'est jamais négligeable devant $O(n^{-(k+1)/2})$.

6.2 Deuxième type de développements

Avec les mêmes notations et les mêmes concepts que dans la section précédente, nous obtenons aussi des résultats asymptotiques pour $\mathbb{P}_a(S_n \geq s_n)$. En effet, la proposition 3.12 donne

$$\mathbb{P}_a(S_n \geq s_n) = \mathrm{e}^{-n\,\Lambda^\star(v)-t\,(s_n-nv)}\,G(t)_a\,D'_n,$$
$$\text{où } D'_n := \mathbb{E}_a^{(t)}\left[\mathrm{e}^{-t\,(S_n-s_n)}\,J_{X_n}\,;\,S_n \geq s_n\right], \quad J_a = 1/G(t)_a.$$

Fixons un entier k. L'hypothèse (6.1) passe à la chaîne de Markov twistée, donc nous pouvons appliquer le théorème 5.4. Ainsi, il existe une constante $C_k > 0$, telle que pour tout $x \in \mathbb{Z}^d$,

$$\left|\mathbb{E}_a^{(t)}[J_{X_n}\,;\,S_n = x] - n^{-d/2}\sum_{j=0}^{k} n^{-j/2}\psi_a^j\left(\frac{x-nv}{\sqrt{n}}\right)\right| \leq \frac{C_k}{n^{(d+k+1)/2}}.$$

Sommons ceci sur $x = y + s_n$, avec $y \in \mathbb{Z}^d$, $y \geq 0$, après avoir multiplié le terme en x par $\mathrm{e}^{-t\,y}$. En posant

$$U_n^j := \sum_{y\geq 0} \mathrm{e}^{-t\,y}\psi_a^j\left(\frac{s_n+y-nv}{\sqrt{n}}\right),$$

il vient

$$n^{d/2}\,D'_n = \sum_{j=0}^{k} n^{-j/2}\,U_n^j + C_k\,n^{-(k+1)/2}\,I_9(n),$$

avec

$$|I_9(n)| \leq D(t), \qquad D(t) := \sum_{y\geq 0}\mathrm{e}^{-t\,y}.$$

Ce qui donne le théorème suivant.

Théorème 6.4. *Quand $n \to \infty$,*

$$\mathbb{P}_a(S_n \geq s_n) = n^{-d/2}\mathrm{e}^{-t\,(s_n-nv)-n\Lambda^\star(v)}\left(\sum_{j=0}^{k} n^{-j/2}U_n^j + O(n^{-(k+1)/2})\right).$$

Comme dans la section précédente, on obtient un équivalent en appliquant ce théorème avec $k = 0$.

Corollaire 6.5. *Avec les notations du corollaire 6.2, quand $n \to +\infty$,*

$$\mathbb{P}_a(S_n \geq s_n) \sim C'(a,\,v)\,n^{-d/2}\,\mathrm{e}^{-t\,(s_n-nv)-n\Lambda^\star(v)},$$

où $C'(a,\,v) := C(a,\,v)/\eta(v)$, avec $\eta(v) = \prod_{j=1}^{d}(1-\mathrm{e}^{-t_j})$.

Démonstration. Comme dans la démonstration du corollaire 6.2, avec $k = 0$, la seule fonction ψ_a^j qui apparaît dans le développement est $\psi_a^0(\cdot) = (\pi(t)\, J)\, \varphi_{\Gamma(t)}(\cdot)$. Donc,

$$U_n^0 := (\pi(t)\, J)\, D_n'' \quad \text{avec}$$
$$D_n'' := \sum_{y \geq 0} \mathrm{e}^{-t\, y} d_n(y), \qquad d_n(y) := \varphi_{\Gamma(t)} \left(\frac{y + s_n - nv}{\sqrt{n}} \right).$$

Pour tout y fixé, $0 \leq d_n(y) \leq \langle \Gamma(t) \rangle$ et $d_n(y) \to \langle \Gamma(t) \rangle$. Avec le théorème de convergence dominée pour les séries, il vient $D_n'' \to D(t)\, \langle \Gamma(t) \rangle$. Comme $D(t)$ est la somme d'une série géométrique, $D(t) = 1/\eta(t)$. Ce qui démontre le résultat voulu. □

Remarque 6.3. Contrairement aux résultats pour $\mathbb{P}_a(S_n = s_n)$ dans les corollaires 6.2 et 6.3, la vitesse de convergence vers le régime décrit ci-dessus dépend de la valeur de t. En effet, chaque $d_n(y) \to \langle \Gamma(t) \rangle$ et plus y est loin de l'origine, plus cette convergence est lente. De plus, quand t est proche de l'origine, beaucoup de termes $d_n(y)$ dans D_n'' ont un effet non négligeable sur la somme, en particulier des termes pour lesquels y est loin de l'origine. Donc, la convergence vers le régime asymptotique du corollaire 6.5 est nettement plus lente quand $t = \Xi(v)$ est proche de l'origine, autrement dit quand v est proche de m. Plus précisément, regardons le comportement de

$$V_n := \left(\lim_{n \to \infty} U_n^0 \right) - U_n^0.$$

Soit K un majorant de toutes les dérivées partielles d'ordre 1 de $\varphi_{\Gamma(t)}$. Alors, les théorèmes des accroissements finis donnent

$$0 \leq V_n \leq n^{-1/2} K\, (\pi(t)\, J) \sum_{y \geq 0} e^{-t\, y} \left\| y + o(1) \right\|.$$

Et cette dernière somme augmente fortement quand t est proche de l'origine.

Remarque 6.4. En comparant les corollaires 6.2 et 6.5, nous constatons que $\mathbb{P}_a(S_n = s_n)$ et $\mathbb{P}_a(S_n \geq s_n)$ sont équivalents à une constante près. Le comportement local asymptotique de S_n au point s_n est donc du même ordre que celui de S_n sur $\{x \in \mathbb{Z}^d\,;\, x \geq s_n\}$.

Chapitre 7

Simulations

Sommaire

Dans ce chapitre, nous estimons les probabilités de grandes déviations $\mathbb{P}_\mu(S_n \in B_n)$, où $(B_n)_{n\in\mathbb{N}}$ est une suite de boréliens fixée et μ la loi initiale de la chaîne, par le biais de simulations pour vérifier nos résultats. Nous supposons que B_n/n ne contient pas la dérive de S_n sous $\mathbb{P}_\mu$. Alors, nous savons bien qu'il n'est pas possible d'estimer directement la probabilité $\mathbb{P}_\mu(S_n \in B_n)$, puisque cette probabilité est exponentiellement petite. Cependant, d'après la proposition 3.12, il suffit d'estimer

$$D_n := \mathbb{E}_\mu^{(t)} \left(\mathrm{e}^{-t\,(S_n - nv)}\, J_{X_n}\, ;\, S_n \in B_n \right),$$

pour un certain v dans l'intérieur de l'enveloppement convexe $\mathcal{E}$ de $f(\Sigma)$ et $t = \Xi(v) \in \mathbb{R}^d$. Ce qui va nous permettre de vérifier les résultats énoncés dans le chapitre 6, en particulier les corollaires 6.2 et 6.5., où v est la dérive de S_n sous $\mathbb{P}^{(t)}$. Deux types de suites (B_n) nous intéressent : $B_n = \{s_n\}$ ou bien $B_n = \{x \in \mathbb{R}^d\, ;\, x \geq s_n\}$, où (s_n) est une suite de $\mathbb{Z}^d$ fixée, telle que $s_n = nv + O(1)$ quand n tend vers l'infini.

Dans toute la suite du chapitre, seule la chaîne twistée intervient. On peut utiliser les notations suivantes : $\mathbb{P} = \mathbb{P}^{(t)}$ est la mesure de probabilité sous laquelle X est la chaîne twistée, π sa distribution stationnaire, v la dérive de S_n, et Γ la matrice de covariance asymptotique de S_n, ie

$$\mathbb{E}_\mu(\bar{S}_n\, \bar{S}_n^{\,t}) \sim n\Gamma, \quad \text{où } \bar{S}_n = S_n - nv.$$

```
"figure5" <-
  function(n_max,N)
{
  # initialisation
  x <- floor( 7 * runif(N) )
  sn1 <- seq( length=N, from=0, by=0 )
  sn2 <- seq( length=N, from=0, by=0 )
  sn3 <- seq( length=N, from=0, by=0 )
  retour <- list(resultat=c(), ecarttype=c())
  # calcul des valeurs successives de S_n
  for (n in 1:n_max)
    {
      # calcul des nouvelles coordonnées de S_n
      sn1 <- sn1 + (x==0) - (x==1)
      sn2 <- sn2 + (x==2) - (x==3)
      sn3 <- sn3 + (x==4) - (x==5)
      # stockage du résultat
      retour$resultat <- c( retour$resultat,
                           n^(3/2)*mean( exp(-.3*sn1-.4*sn2-.3*sn3)*
                                        (sn1>=0)*(sn2>=0)*(sn3>=0) ) )
      retour$ecarttype <- c( retour$ecarttype,
                            n^(3/2)*sd( exp(-.3*sn1-.4*sn2-.3*sn3)*
                                       (sn1>=0)*(sn2>=0)*(sn3>=0) ) )
      # changement d'état de la chaîne de Markov
      x <- ( x + ( runif(N)>0.7 ) ) %% 7
    }
  retour
}
```

FIG. 7.1 – Le code qui a permis de faire la figure 5. Le paramètre `N` est la taille de l'échantillon et `n_max` la valeur maximale de n.

7.1 Buts et méthodes

Nous voulons vérifier les prédictions suivantes :
- le comportement de $\mathbb{P}_\mu(S_n \in B_n)$ dépend de la dimension à travers le facteur $n^{-d/2}$,
- le facteur $\mathrm{e}^{-t(s_n - nv)}$ peut conduire à des oscillations périodiques,
- les deux corollaires 6.3 et 6.5 sont vrais quelque soit l'état initial et, par combinaison linéaire, quelque soit la mesure initiale,
- les constantes $C(a, v)$ et $C'(a, v)$ sont correctes,
- la vitesse de convergence vers le régime asymptotique dans le corollaire 6.5 dépend bien de t.

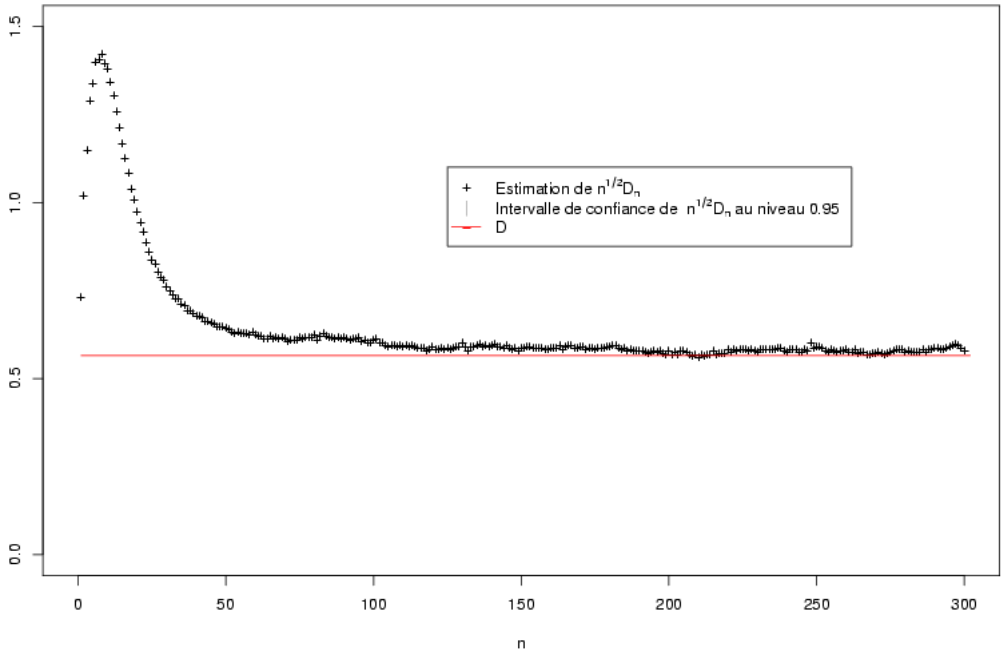

FIG. 7.2 – Soit $f = \mathbf{1}_0 - \mathbf{1}_1$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 7 où $\{U_n\}$ sont i.i.d. de loi $0.7\delta_0 + 0.3\delta_1$, $S_n = \sum_1^n f(X_i)$, $t = 2$, et $D_n = \mathbb{E}_\pi(\mathrm{e}^{-t\,S_n}\,;\, S_n \geq 0)$. Alors $n^{1/2}\,D_n$ converge vers $D \approx 0,5650774$.

Concernant la partie technique des simulations, nous avons utilisé l'environnement et le langage *R* [51]. Nous donne un exemple de code dans la figure 7.1. Nous avons employé différents types de générateurs de nombres pseudo-aléatoires, qui donnent tous les mêmes résultats. Nous présentons ici ceux que nous avons obtenu avec Mersenne-Twister, générateur dont la période est $2^{19937} - 1$, voir Matsumoto et Nishimura [41]. Nous estimons $n^{d/2}D_n$ en réalisant un échantillon de taille N de chaînes de Markov de loi $Q^{(t)}$, ce qui nous donne N suites de variables aléatoires $(S_n(j))_{n\in\mathbb{N}}$, $1 \leq j \leq N$. Notre estimateur de $n^{d/2}\,D_n$ est

$$Z_n(N) = n^{d/2}\,N^{-1}\sum_{j=1}^{N} g(S_n(j),\,X_n(j)), \quad \text{où } g(y,\,a) = \mathrm{e}^{-t\,(y-nx)}\,\mathbf{1}_{B_n}(y)\,J_a.$$

Comme les $g(S_n(j),\,X_n(j))$, $1 \leq j \leq N$ sont des variables aléatoires i.i.d., la variance de $Z_n(N)$ vérifie

$$\text{var}\,(Z_n(N)) = n^d\,N^{-1}\,\text{var}\,g(S_n,\,X_n).$$

Les lemmes suivants donnent le comportement asymptotique de cette variance quand $n \to \infty$, N étant fixé.

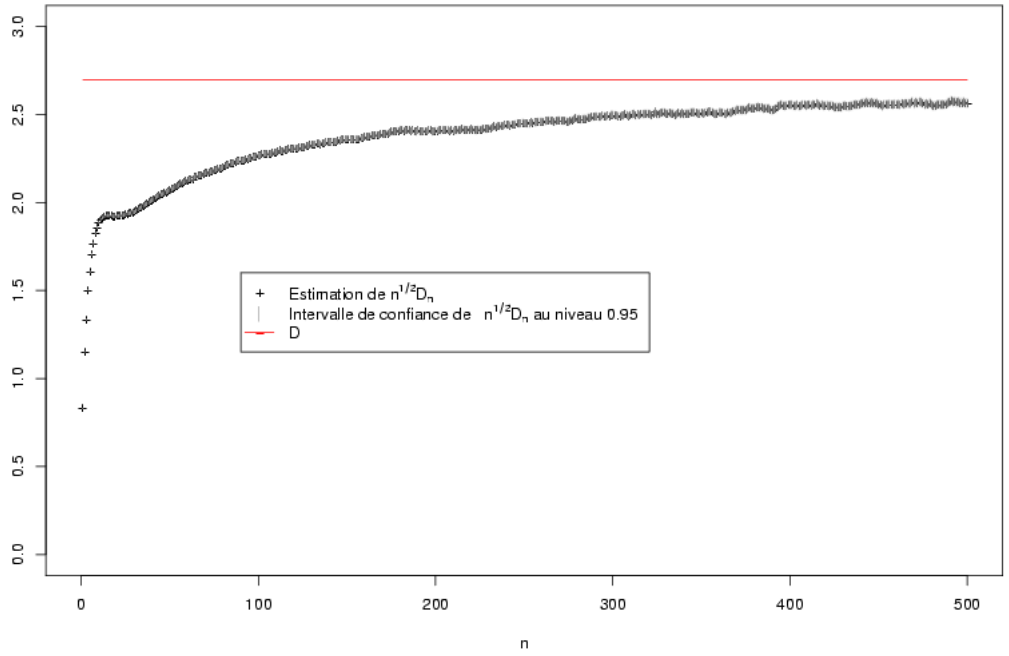

FIG. 7.3 – Soit $f = \mathbf{1}_0 - \mathbf{1}_1$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 7 où $\{U_n\}$ sont i.i.d. de loi $0.7\delta_0 + 0.3\delta_1$, $S_n = \sum_1^n f(X_i)$, $t = 0, 2$, et $D_n = \mathbb{E}_\pi(\mathrm{e}^{-t\,S_n}\,;\, S_n \geq 0)$. Alors $n^{1/2}\,D_n$ converge vers $D \approx 2.6954$.

Lemme 7.1. *Si* $\{S_n \in B_n\} = \{S_n = s_n\}$, $s_n = nv + O(1)$, *alors*

$$\operatorname{var} Z_n(N) = n^{d/2}\,N^{-1}\big(C_1 + O(n^{-1/2})\big)\,\mathrm{e}^{-2t\,(s_n - nv)},$$

où

$$C_1 := \langle\Gamma(t)\rangle\,\big((\pi(t)J^2) - \langle\Gamma(t)\rangle\,(\pi(t)J)^2\big).$$

Démonstration. Nous devons étudier

$$\mathbb{E}_\mu\left(g(S_n,\,X_n)^2\right) = \mathbb{E}_\mu\left(\mathrm{e}^{-2t\,(S_n - nx)}\,J_{X_n}^2\,;\,S_n = s_n\right).$$

Comme pour le corollaire 6.2, le théorème 5.4 avec $k = 0$ montre que

$$\mathbb{E}_\mu\left(g(S_n,\,X_{n+1})^2\right) = n^{-d/2}(K_1 + O(n^{-1/2}))\,\mathrm{e}^{-2t\,(s_n - nv)}$$

avec $K_1 = \langle\Gamma(t)\rangle\,(\pi(t)J^2)$. En élevant au carré le résultat du corollaire 6.2, on obtient

$$\left(\mathbb{E}_\mu(g(S_n,\,X_n))\right)^2 = n^{-d/2}(K_2 + O(n^{-1/2}))\,\mathrm{e}^{-2t\,(s_n - nv)}$$

avec $K_2 = \langle\Gamma(t)\rangle^2\,(\pi(t)J)^2$. D'où le résultat. □

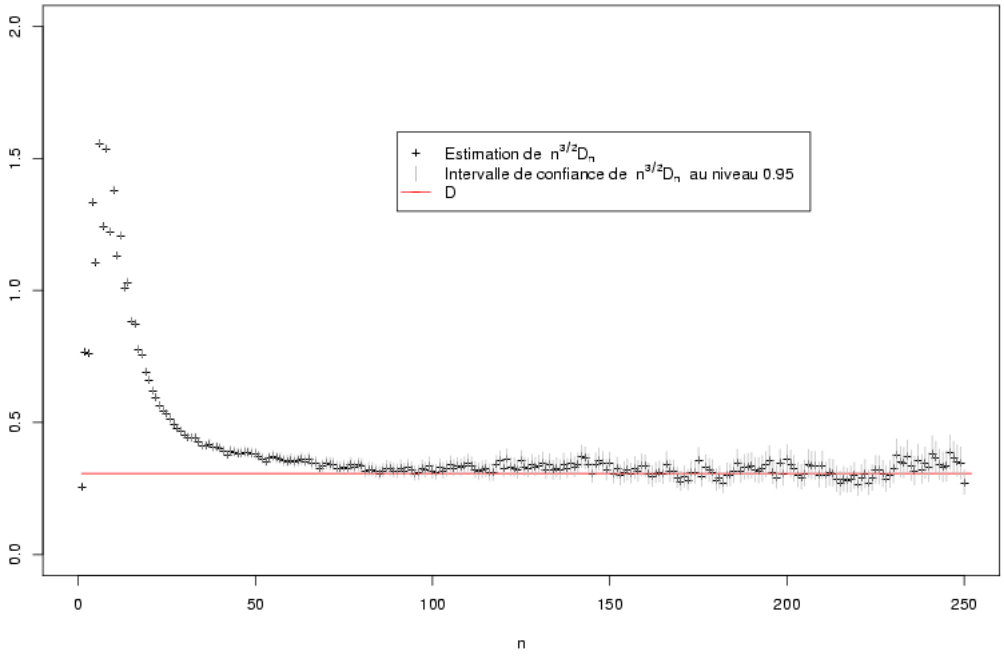

FIG. 7.4 – Soit $f = (\mathbf{1}_0 - \mathbf{1}_1, \mathbf{1}_2 - \mathbf{1}_3, \mathbf{1}_4 - \mathbf{1}_5)$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 7 où $\{U_n\}$ sont i.i.d. de loi $0.7\delta_0 + 0.3\delta_1$, $S_n = \sum_1^n f(X_i)$, $t = (1, 3, 1)$, et $D_n = \mathbb{E}_\pi(\mathrm{e}^{-t\,S_n}\,;\, S_n \geq 0)$. Alors $n^{3/2}\,D_n$ converge vers $D \approx 0,3072178$.

Lemme 7.2. *Si* $\{S_n \in B_n\} = \{S_n \geq s_n\}$, $s_n = nv + O(1)$, *alors*

$$\operatorname{var} Z_n(N) = n^{d/2}\,N^{-1}\big(C_2 + o(1)\big)\,\mathrm{e}^{-2t(s_n - nv)},$$

où

$$C_2 := \langle \Gamma(t) \rangle \left(\frac{\pi(t)\,J^2}{\eta(2t)} - \langle \Gamma(t) \rangle \frac{(\pi(t)\,J)^2}{\eta(t)^2} \right)$$

et $\eta(t) = \prod_{j=1}^d (1 - \mathrm{e}^{-t_j})$.

Démonstration. Nous devons étudier

$$\mathbb{E}_\mu\left(g(S_n,\, X_n)^2\right) = \mathbb{E}_\mu\left(\mathrm{e}^{-2t\,(S_n - nx)}\,J_{X_n}^2\,;\, S_n \geq s_n\right).$$

On utilise le théorème de convergence dominé pour les séries comme dans la démonstration du corollaire 6.5. Il vient

$$\mathbb{E}_\mu\left(g(S_n,\, X_n)^2\right) = n^{-d/2}(K_1 + o(1))\,\mathrm{e}^{-2t\,(s_n - nv)}$$

avec $K_1 = \langle \Gamma(t) \rangle\,(\pi(t)J^2)/\eta(2t)$. Il suffit d'élever au carré le résultat du corollaire 6.5 pour obtenir un équivalent de $(\mathbb{E}_\mu g(S_n,\, X_n))^2$. □

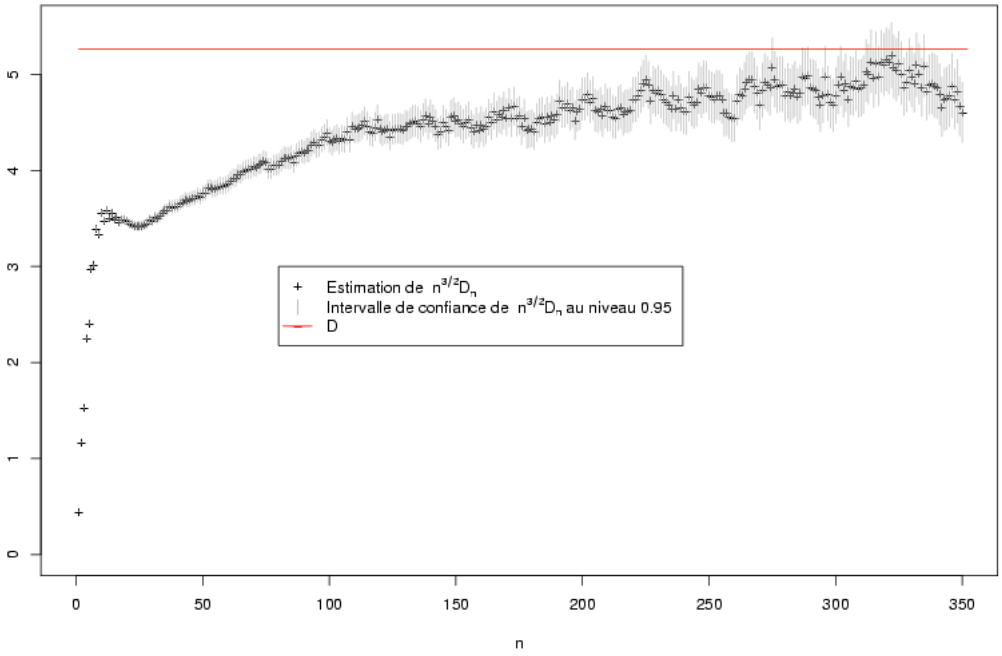

FIG. 7.5 – Soit $f = (\mathbf{1}_0 - \mathbf{1}_1, \mathbf{1}_2 - \mathbf{1}_3, \mathbf{1}_4 - \mathbf{1}_5)$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 7 où $\{U_n\}$ sont i.i.d. de loi $0.7\delta_0 + 0.3\delta_1$, $S_n = \sum_1^n f(X_i)$, $t = (0,3;\, 0,4;\, 0,3)$, et $D_n = \mathbb{E}_\pi(\mathrm{e}^{-t\,S_n}\,;\, S_n \geq 0)$. Alors $n^{3/2}\,D_n$ converge vers $D \approx 5,267$.

Remarquons que ces deux lemmes montrent que la variance de $Z_n(N)$ augmente avec n en $n^{d/2}$, mais diminue avec la taille de l'échantillon N en N^{-1}. Nous devons donc trouver un compromis entre la taille de l'échantillon et la valeur maximale de n que l'on veut atteindre dans les simulations pour que l'intervalle de confiance garde une taille raisonnable par rapport à la quantité estimée. Nous utilisons un estimateur sans biais de la variance de l'échantillon pour construire un intervalle de confiance de niveau 95 %.

Les valeurs des constantes $C(a,\, v)$ et $C'(a,\, v)$ sont simples à calculer puisque nous connaissons la loi de la chaîne de Markov sous $Q^{(t)}$ dans chaque simulation. En fait, nous calculons $\Gamma(t)$ et utilisons les corollaires 6.2 et 6.5 pour obtenir des valeurs numériques de ces limites.

7.2 Discussion sur les simulations

Nous avons fait des simulations avec différentes valeurs de d pour vérifier que le comportement de D_n dépend de la dimension à travers le facteur $n^{-d/2}$.

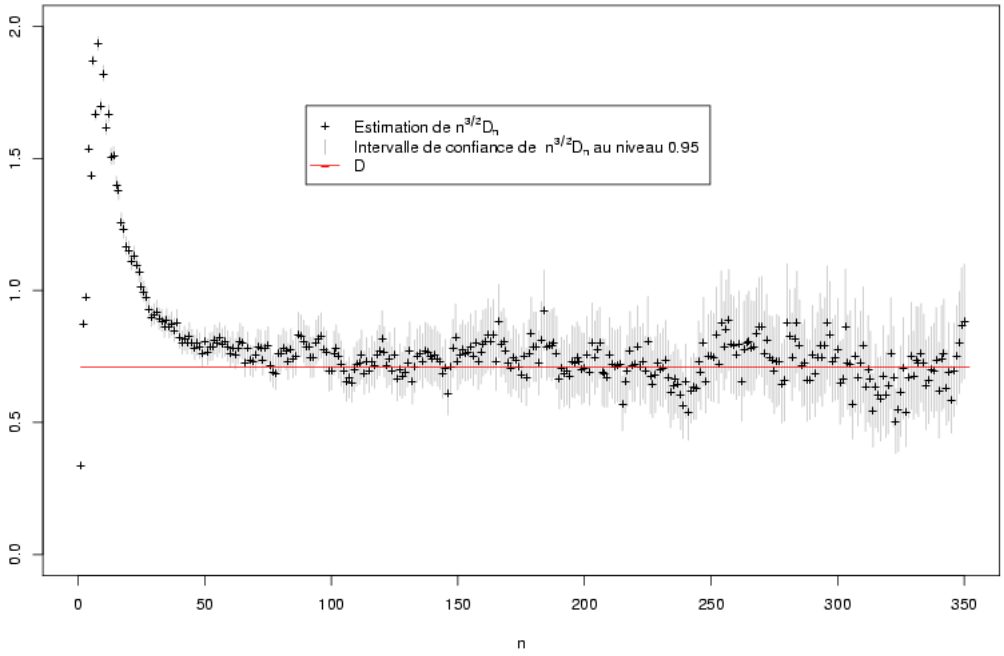

FIG. 7.6 – Soit $f = (\mathbf{1}_0 - \mathbf{1}_1, \mathbf{1}_2 - \mathbf{1}_3, \mathbf{1}_4 - \mathbf{1}_5)$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 7 où $\{U_n\}$ sont i.i.d. de loi $0.7\delta_0 + 0.3\delta_1$, $S_n = \sum_1^n f(X_i)$, $t = (0,8;\, 0,7;\, 0,9)$, et $D_n = \mathbb{E}_0(\mathrm{e}^{-t\,S_n}\,;\, S_n \geq 0)$. Alors $n^{3/2}\,D_n$ converge vers $D \approx 0,7090541$.

Les différentes figures montrent des simulations pour $d = 1$,2 ou 3. Malgré un échantillon de taille $N = 400^2$, nous voyons que l'intervalle de confiance pour $d = 3$ dans les figures 7.4 et 7.5 augmente beaucoup plus avec n que pour $d = 1$ (figures 7.2 et 7.3). C'est pour cette raison que nous sommes limités à $d = 3$.

Ces simulations nous permettent de vérifier que la constante $C'(a,\, v)$ donnée dans le corollaire 6.5 est correcte. En effet, nous observons que $n^{d/2}D_n$ converge bien vers la limite D donnée par ce résultat dans les figures 7.2, 7.3, 7.4 et 7.5, où le terme $\mathrm{e}^{-t\,(s_n - nv)}$ n'intervient pas. Nous pouvons également vérifier que la constante $C(a,\, v)$ donnée dans les corollaires 6.2 et 6.3 est correcte. Dans les figures 7.10, 7.11 et 7.12, $n^{d/2}D_n$ converge bien vers la limite prédite ou les valeurs d'adhérence prédites.

Lorsque le terme $\mathrm{e}^{-t\,(s_n - nv)}$ intervient, nous observons de plus les oscillations qu'il permet de prévoir, voir les figures 7.7, 7.9 et 7.12. De plus, on constate que plus ce facteur est élevé, plus l'intervalle de confiance est grand, même pour des valeurs de n successives. Ceci s'explique par le facteur $\mathrm{e}^{-2t\,s_n - nv)}$ dans la variance de $g(S_n,\, X_{n+1})$ donné par le lemme 7.2. En effet, dans les figures 7.2, 7.3, 7.4 et 7.5, $v = 0$ est dans $\mathbb{Z}^d$, $s_n = nv$ est donc $\mathbb{Z}^d$, et $\mathrm{e}^{-t\,(s_n - nv)} = 1$

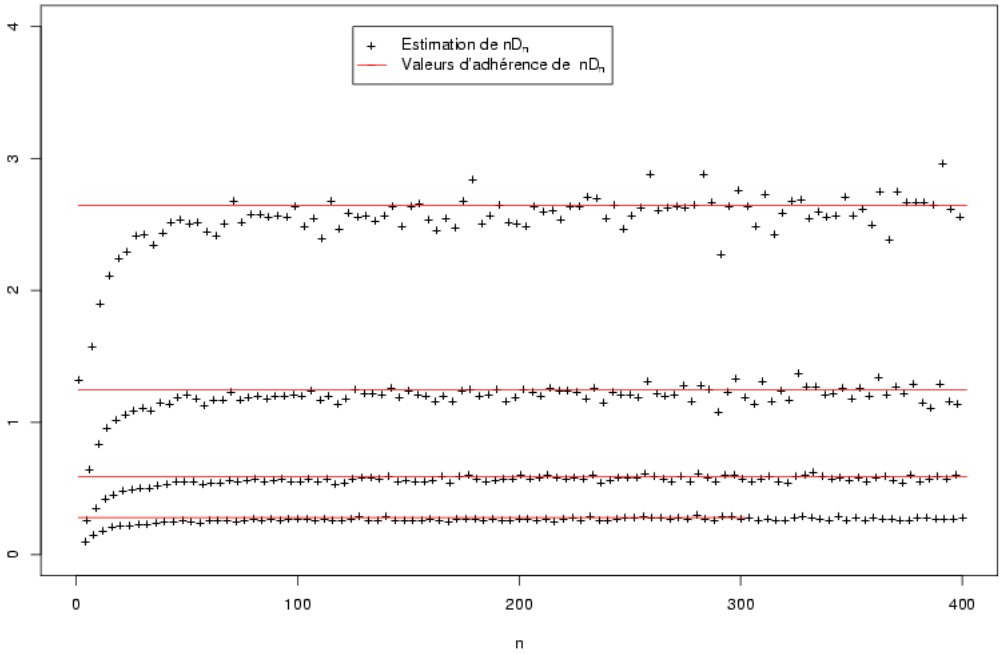

FIG. 7.7 – Soit $f = (\mathbf{1}_0, \mathbf{1}_1)$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 4 où $\{U_n\}$ sont i.i.d. de loi $0.8\delta_0 + 0.1\delta_1 + 0.1\delta_2$, $S_n = \sum_1^n f(X_i)$, $t = (1, 2)$, $v = (1/4, 1/4)$, et $D_n = \mathbb{E}_\pi(\mathrm{e}^{-t\,(S_n - nv)}\,;\, S_n \geq \lfloor nv \rfloor)$. Alors $u(n) = n\,D_n$ oscille périodiquement. En effet, $u(4n+i)$ converge vers $D\,\mathrm{e}^{\beta_i}$ quand $n \to +\infty$, avec $D \approx 0,2784627$, $\beta_0 = 0$, $\beta_1 = 3/4$, $\beta_2 = 3/2$ et $\beta_3 = 9/4$.

pour tout n. En revanche, dans les figures 7.7, 7.9 et 7.12, le facteur $\mathrm{e}^{-t\,(s_n - nv)}$ conduit à des oscillations. Pour comprendre ceci, notons que $\{S_n \geq n\,v\} = \{S_n \geq \lceil nv \rceil\}$ où, $\lceil nv \rceil$ est définie, coordonnées par coordonnées, comme la partie entière supérieure. De plus, quand les coordonnées de v sont rationnelles, la suite $(\lceil nv \rceil - nv)_{n \in \mathbb{N}}$ est périodique, de période 6 dans le cas de la figure 7.9 et l'on retrouve bien cette période dans les simulations. Dans les figures 7.7 et 7.12, nous avons utilisé la partie entière inférieure, c'est-à-dire $s_n = \lfloor nv \rfloor$. Nous obtenons des oscillations similaires, de période 4, qui est exactement la période de $\lfloor nv \rfloor$, c'est-à-dire la période de $s_n - nv$.

De plus, en comparant les figures 7.2 et 7.3 d'une part et les figures 7.4 et 7.5, nous observons que la convergence de $n^{d/2} D_n$ vers le régime asymptotique est plus lent quand t est proche de l'origine. Ceci correspond bien à ce que nous avons noté dans la remarque 6.3 à la suite de la démonstration du corollaire 6.5 dans le chapitre 6. Ce phénomène n'apparaît pas pour $\mathbb{P}^{(t)}(S_n = s_n)$. En effet, dans ce cas, t n'apparaît plus que dans le noyau de transition $Q^{(t)}$ que nous fixons et dans le terme d'oscillations $\mathrm{e}^{-t\,(s_n - nv)}$.

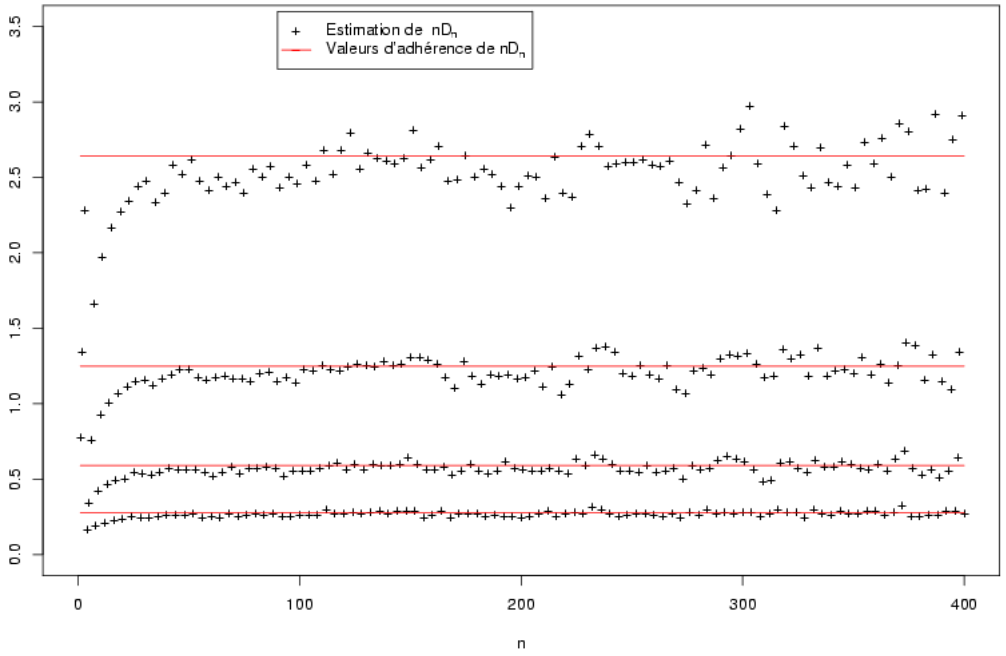

FIG. 7.8 – Soit $f = (\mathbf{1}_0, \mathbf{1}_1)$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 4 où $\{U_n\}$ sont i.i.d. de loi $0.8\delta_0 + 0.1\delta_1 + 0.1\delta_2$, $S_n = \sum_1^n f(X_i)$, $t = (1, 2)$, $v = (1/4, 1/4)$, et $D_n = \mathbb{E}_0(\mathrm{e}^{-t\,(S_n - nv)}\,;\, S_n \geq \lfloor nv \rfloor)$. Alors $u(n) = n\,D_n$ oscille périodiquement. En effet, $u(4n+i)$ converge vers $D\,\mathrm{e}^{\beta_i}$ quand $n \to +\infty$, avec $D \approx 0,2784627$, $\beta_0 = 0$, $\beta_1 = 3/4$, $\beta_2 = 3/2$ et $\beta_3 = 9/4$.

De plus, l'estimation obtenue est correcte quelque soit la mesure initiale de la chaîne de Markov. Par exemple, dans la figure 7.5 la mesure initiale de la chaîne de Markov est la distribution stationnaire π, alors que dans la figure 7.6, la mesure initiale ne charge que l'état 0. De même, la seule différence entre les figures 7.7 et 7.8 est la mesure initiale : dans la première figure, celle-ci est la distribution stationnaire π et dans la seconde, celle-ci ne charge que l'état 0. Dans les figures 7.13 et 7.14, la mesure initiale de la chaîne de Markov est la distribution uniforme, qui n'est pas la distribution stationnaire de la chaîne.

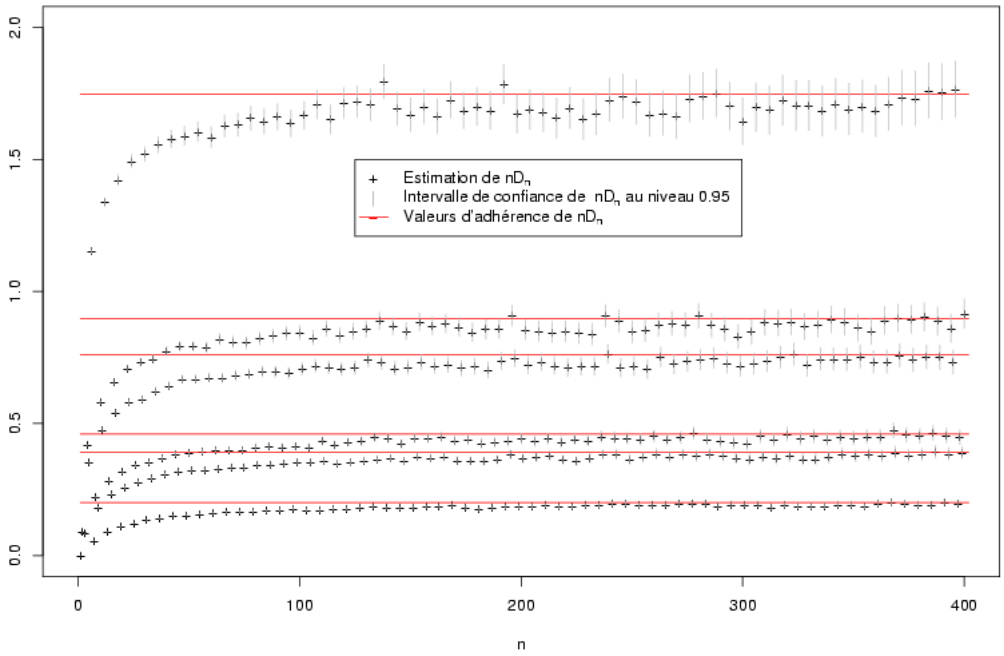

FIG. 7.9 – Soit $f = (\mathbf{1}_0, \mathbf{1}_1)$, $\{X_n\}$ i.i.d. de loi $\delta_0/2+(\delta_1+\delta_2+\delta_3)/6$, $S_n = \sum_1^n f(X_i)$, $t = (1, 2)$, $v = (1/2, 1/6)$, et $D_n = \mathbb{E}_\pi(\mathrm{e}^{-t\,(S_n - nv)}\,;\, S_n \geq nv)$. Alors $u(n) = n\,D_n$ oscille périodiquement. En effet, $u(6n + i)$ converge vers $D\,\mathrm{e}^{\beta_i}$ quand $n \to +\infty$, avec $D \approx 1,747124$, $\beta_0 = 0$, $\beta_1 = -13/6$, $\beta_2 = -4/3$, $\beta_3 = -3/2$, $\beta_4 = -2/3$, et $\beta_5 = -5/6$.

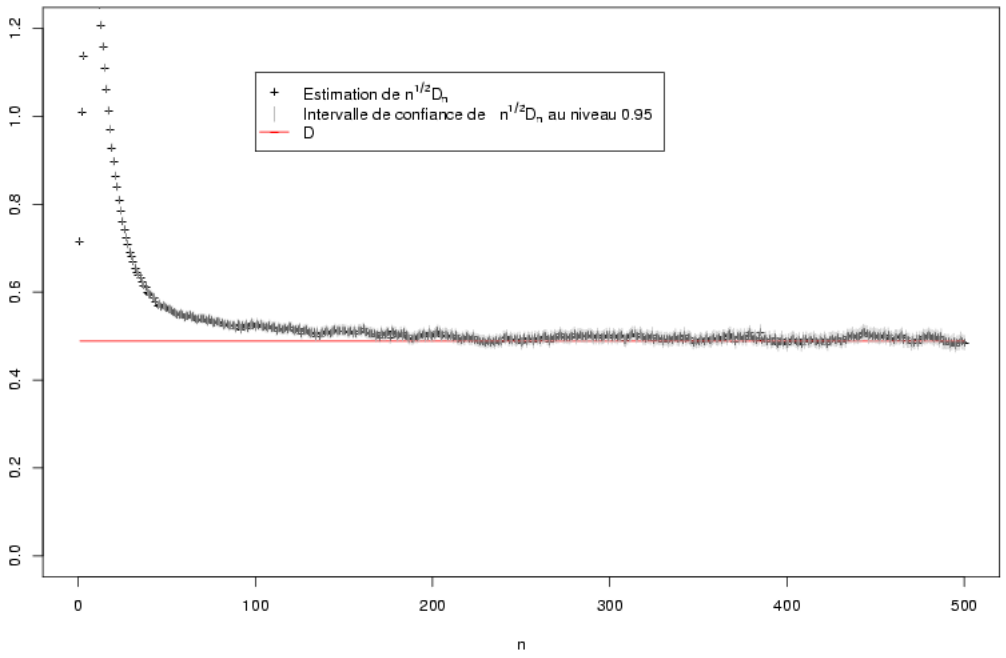

FIG. 7.10 – Soit $f = \mathbf{1}_0 - \mathbf{1}_1$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 7 où $\{U_n\}$ sont i.i.d. de loi $0.7\delta_0 + 0.3\delta_1$, $S_n = \sum_1^n f(X_i)$, et $D_n = \mathbb{P}_\pi(S_n = 0)$. Alors $n^{1/2}\, D_n$ converge vers $D \approx 0,488602$.

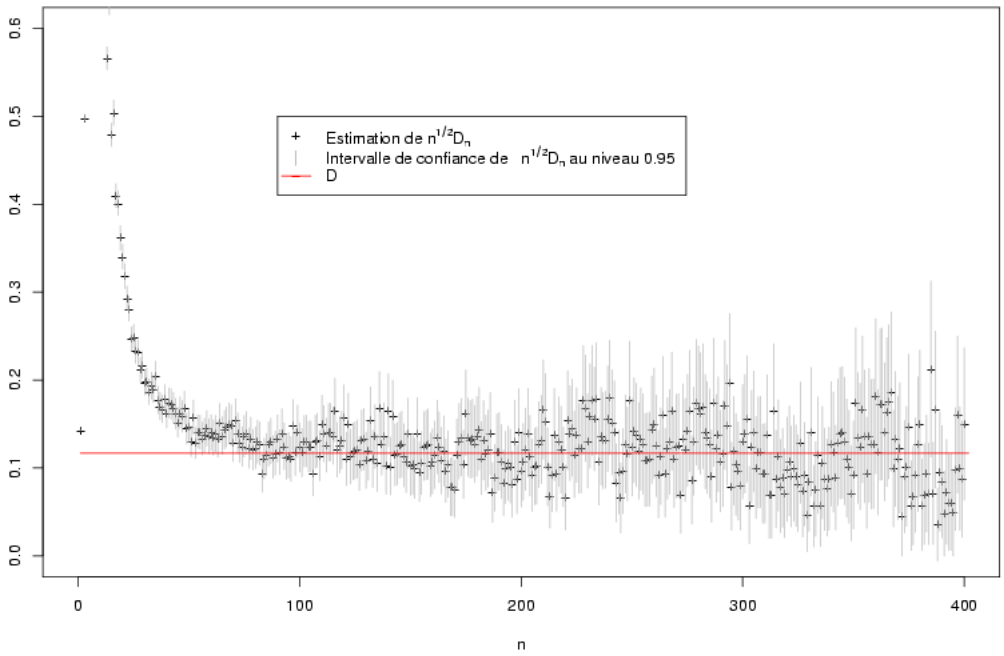

FIG. 7.11 – Soit $f = (\mathbf{1}_0 - \mathbf{1}_1, \mathbf{1}_2 - \mathbf{1}_3, \mathbf{1}_4 - \mathbf{1}_5)$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 7 où $\{U_n\}$ sont i.i.d. de loi $0.7\delta_0 + 0.3\delta_1$, $S_n = \sum_1^n f(X_i)$, $t = (1, 3, 1)$, et $D_n = \mathbb{P}_\pi(S_n = 0)$. Alors $n^{3/2}\, D_n$ converge vers $D \approx 0.1166453$.

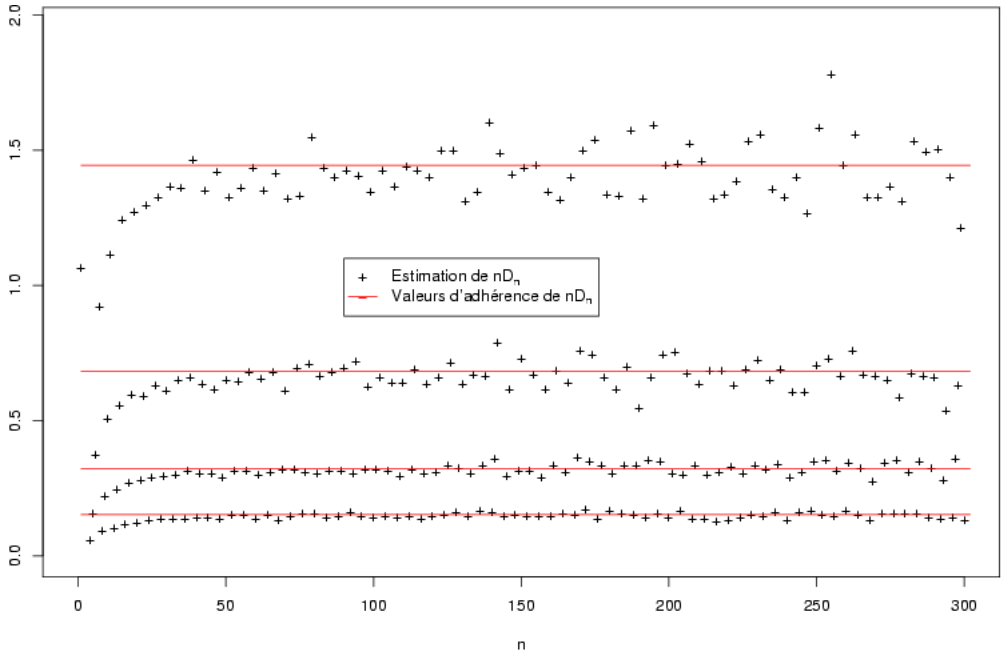

FIG. 7.12 – Soit $f = (\mathbf{1}_0, \mathbf{1}_1)$, $\{X_n\}$ est la chaîne de Markov définie par $X_{n+1} = X_n + U_n$ modulo 4 où $\{U_n\}$ sont i.i.d. de loi $0.8\delta_0 + 0.1\delta_1 + 0.1\delta_2$, $S_n = \sum_1^n f(X_i)$, $t = (1, 2)$, $v = (1/4, 1/4)$, et $D_n = \mathbb{E}_\pi(\mathrm{e}^{-t\,(S_n - nv)}\,;\, S_n = \lfloor nv \rfloor)$. Alors $u(n) = n\,D_n$ oscille périodiquement. En effet, $u(4n+i)$ converge vers $D\,\mathrm{e}^{\beta_i}$ quand $n \to +\infty$, avec $D \approx 0,1522$, $\beta_0 = 0$, $\beta_1 = 3/4$, $\beta_2 = 3/2$ et $\beta_3 = 9/4$.

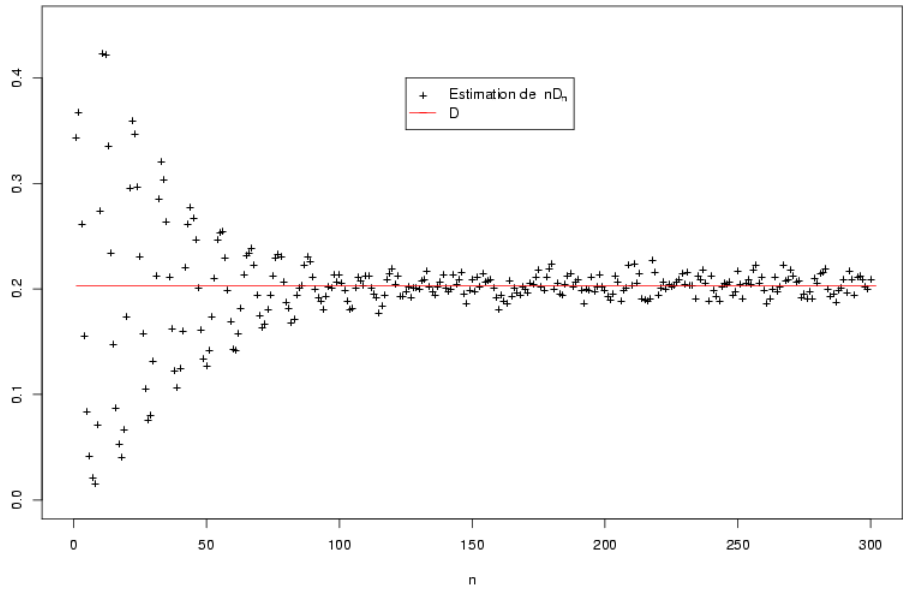

FIG. 7.13 – Soit $f = (\mathbf{1}_0 - 6\mathbf{1}_1, 2\mathbf{1}_1 - \mathbf{1}_2)$, $\{X_n\}$ est la chaîne de Markov sur $\{0, 1, 2, 3\}$ de noyau de transition Q dont les termes non nuls sont $Q(0; 0) = 0,8$, $Q(0; 1) = Q(0;, 2) = 0,1$, $Q(1; 1) = Q(2; 2) = Q(3; 3) = 0,4$ et $Q(1; 2) = Q(2; 3) = Q(3; 0) = 0,6$. $S_n = \sum_1^n f(X_i)$, $t = (2, 1)$, μ la distrubition uniforme sur $\{0, 1, 2, 3\}$ et $D_n = \mathbb{E}_\mu(\mathrm{e}^{-t\,S_n}\,;\, S_n \geq 0)$. Alors $n\,D_n$ converge vers $D \approx 0,203$.

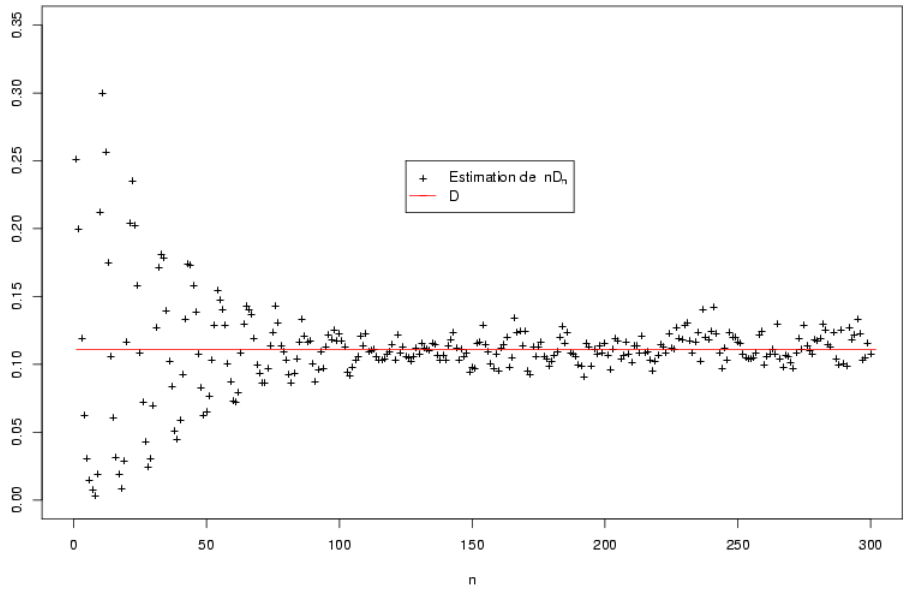

FIG. 7.14 – Soit $f = (\mathbf{1}_0 - 6\mathbf{1}_1, 2\mathbf{1}_1 - \mathbf{1}_2)$, $\{X_n\}$ est la chaîne de Markov sur $\{0, 1, 2, 3\}$ de noyau de transition Q dont les termes non nuls sont $Q(0; 0) = 0,8$, $Q(0; 1) = Q(0;, 2) = 0,1$, $Q(1; 1) = Q(2; 2) = Q(3; 3) = 0,4$ et $Q(1; 2) = Q(2; 3) = Q(3; 0) = 0,6$. $S_n = \sum_1^n f(X_i)$, μ la distrubition uniforme sur $\{0, 1, 2, 3\}$, et $D_n = \mathbb{P}_\mu(S_n = 0)$. Alors $n\,D_n$ converge vers $D \approx 0,111$.

Chapitre 8

Applications à l'étude de séquences biologiques

Sommaire

Les résultats présentés dans le chapitre 6 justifient les méthodes de grandes déviations de recherche de signaux, ayant potentiellement une signification biologique, dans une séquence d'ADN, d'ARN ou de protéine.

On se donne une séquence biologique (d'ADN, d'ARN ou de protéines) et on estime sur cette séquence les paramètres d'un modèle markovien (chaîne de Markov, chaîne de Markov cachée ou modèle r-markovien). Nous cherchons dans cette séquence des mots de longueur ℓ dont la fréquence est mal prédite pour le modèle. Nous les appelons mots de fréquences exceptionnelles ou mots exceptionnels. Plus précisément, nous voulons obtenir une liste de mots de longueur ℓ dont les fréquences sont les plus mal prédites. Le modèle markovien est donc considéré comme du bruit, sur lequel on cherche des signaux. Nous espérons que la mauvaise qualité de la prédiction de la fréquence du mot n'est pas due au hasard, mais imposée par une nécessité biologique.

Nous quantifions l'exceptionnalité d'un mot m par la probabilité que les fréquences de m dans le modèle aléatoire et dans la séquence observée soient voisines. Plus cette probabilité est petite, plus la fréquence est mal prédite et

plus le mot est exceptionnel. Il existe plusieurs méthodes pour estimer cette probabilité. Le calcul direct de cette probabilité par une formule exacte est trop longue quand la longueur n de la séquence est grande. Nous devons remplacer ce calcul exact par une méthode approchée. La méthode que nous utilisons est basée sur les estimées de grandes déviations. Cette méthode a, en outre, le mérite de pouvoir conditionner le modèle aléatoire initial pour que les fréquences observées d'une liste de mots W correspondent à celles prédites par le modèle. Dans toute la suite, on note $\mathcal{A}$ l'alphabet de la séquence. Pour une séquence d'ADN, cet alphabet est $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. Pour les séquences de protéines, l'alphabet $\mathcal{A}$ est l'ensemble des 20 acides aminés.

8.1 Modèles r-markoviens et grandes déviations de la fréquence d'un mot

Soit $r \geq 1$ un entier tel que $r + 1 < \ell$. Supposons que l'on a estimé sur la séquence biologique les paramètres d'un modèle r-markovien. Notons $Y = (Y_j)_{j\in\mathbb{N}}$ une suite de variables aléatoires qui représente ce modèle aléatoire. L'estimation des paramètres nous donne une matrice de transition Π qui vérifie, pour tout $m \in \mathcal{A}^{r+1}$, $j \in \mathbb{N}$,

$$\Pi(m_{1:r},\, m_{r+1}) = \mathbb{P}\left(Y_{j+r+1} = m_{r+1} \,|\, Y_{j+1:j+r} = m_{1:r}\right)$$

Nous supposons dans toute la suite que Π est irréductible et apériodique.

Soit w une liste de mots de longueur ℓ. Définissons S_n comme le nombre d'occurrences de w dans la séquence $Y_{0:n+\ell-1}$ Alors $(S_n)_{n\in\mathbb{Z}}$ est un processus de Markov additif sur $\mathbb{Z}$. En effet, introduisons le processus $X = (X_j)$ défini par $X_j = Y_{j:j+\ell-1}$ pour tout $j \in \mathbb{N}$. Alors $S_n = \sum_{j=0}^{n-1} f(X_j)$, où $f : \Sigma \to \mathbb{Z}$ est la fonction indicatrice de w, avec $\Sigma = \mathcal{A}^\ell$. Et X est une chaîne de Markov sur Σ. Notons Q le noyau de transition de cette chaîne de Markov. Pour tout $m,\, m' \in \Sigma$, il vérifie

$$Q(m,\, m') = \begin{cases} \Pi(m_{\ell-r:\ell},\, m'_\ell) & \text{si } m_{2:\ell} = m'_{1:\ell-1} \\ 0 & \text{sinon.} \end{cases}$$

Comme Π est irréductible et apériodique, Q l'est aussi.

Définition 8.1. *On appelle processus de comptage du mot w la suite de variables aléatoires (S_n) précédemment définie.*

Dans ce cas, nous avons vu au chapitre 3, théorème 3.9, que (S_n) suit un principe de grandes déviations, de vitesse n et de bonne fonction de taux $\Lambda^\star$.

Cette fonction est la transformée de Fenchel-Legendre de la fonction $\Lambda : \mathbb{R} \to \mathbb{R}$, définie par

$$\Lambda^\star(x) = \sup\{t\,x - \Lambda(t)\,;\, t \in \mathbb{R}\}.$$

Le terme $\Lambda(t)$ est la valeur propre dominante du noyau $Q(t)$ dépendant d'un paramètre $t \in \mathbb{R}$ et défini à partir de Q, noyau de transition de X, par

$$Q(t)(a,\, b) = Q(a,\, b)\mathrm{e}^{t\,f(a)}, \quad a, b \in \Sigma.$$

Supposons que w est sur-représenté et cherchons à appliquer les estimations précises de grandes déviations que nous avons démontrées au chapitre 6. Le lemme nous permet d'appliquer les résultats de ce chapitre à notre processus de comptage.

Lemme 8.2. *Le processus de Markov additif (S_n) vérifie : il existe deux mots a et b dans Σ, un vecteur x de $\mathbb{Z}^d$ (en fait, un vecteur de $\mathbb{N}^d$) et un rang n tels que*

$$\mathbb{P}_a(S_n = x\,;\, X_n = b) > 0 \quad \textit{et}$$
$$\mathbb{P}_a(S_n = x + 1\,;\, X_n = b) > 0.$$

Notons $s_n(w) := \lceil n\alpha(w) \rceil$ la partie entière supérieure de $n\alpha(w)$, $\alpha(w)$ étant la fréquence de w observée sur la séquence biologique. Avec les corollaires 6.3 et 6.5, on obtient alors le résultat suivant.

Proposition 8.3. *Pour toute distribution initiale μ de la chaîne de Markov X, il existe une constante K, qui ne dépend que de w et de la mesure initiale μ telle que, quand n tend vers l'infini,*

$$\log \mathbb{P}_\mu(S_n = s_n) = -n\,\Lambda^\star(v) - t\,(s_n - nv) - \frac{1}{2}\log n + K + O(n^{-1/2})$$
$$\log \mathbb{P}_\mu(S_n \geq s_n) = -n\,\Lambda^\star(v) - t\,(s_n - nv) - \frac{1}{2}\log n + K' + o(1)$$

où l'on a posé $v = \alpha(w)$, $t = \Xi(v) > 0$, Ξ étant l'homéomorphisme réciproque de Λ'.

Posons $I(w) := \Lambda^\star(v)$. Le terme dominant de ces développements asymptotiques est $-nI(w)$. Ainsi, plus $I(w)$ est grand, plus la probabilité que la fréquence observée soit voisine de la fréquence observée dans le modèle est faible. Donc, plus $I(w)$ est grand, plus la fréquence de w est mal prédite. Quand w est sous-représenté, on obtient les mêmes résultats avec $\mathbb{P}_\mu(S_n = s_n)$ et $\mathbb{P}_\mu(S_n \leq s_n)$, où $s_n = \lfloor nv \rfloor$ est la partie entière de nv et $v = \alpha(w)$ est la fréquence de w observée.

Dans sa thèse, Nuel [49, partie II] utilise $I(w)$ pour quantifier le caractère exceptionnel de certains mots, et ainsi vérifier leur exceptionnalité. Cependant,

le calcul de $I(w)$ via la transformée de Fenchel-Legendre de la valeur propre dominante de $Q(t)$ est délicat. En effet, il faut maximiser la fonction concave $t \mapsto tv - \Lambda(t)$ sur $\mathbb{R}$ par un algorithme de type descente du gradient. Et $\Lambda(t)$ est la valeur propre d'une matrice de Perron-Frobenius. Il faut donc avoir une estimation numérique de cette valeur propre en chaque point t par laquelle passe l'algorithme de descente du gradient. La méthode de calcul de cette valeur propre dominante, basée sur la méthode d'Arnoldi, est coûteuse en mémoire et en temps lorsqu'on veut l'implémenter, voir Nuel [49, Annexe C]. Nuel compare également les valeurs de $I(w)$ pour tous les mots w de longueur $\ell = 9$ dans les modèles 1-markoviens de différents génomes, voir [49, chapitre 11]. Il obtient ainsi les résultats des figures 1.1, 2.1, 3.1, 4.1, 5.1, 6.1, 7.1, 8.1, 9.1, 10.1 et 11.1.

8.2 Conditionnements successifs du modèle aléatoire

Plutôt que de comparer le caractère exceptionnel de deux mots donnés, et de classer tous les mots de longueur ℓ du plus mal prédit au mieux prédit par un modèle donné, nous proposons une autre méthode pour obtenir une liste de mots exceptionnels plus réduite. Il est important d'obtenir la liste de mots la plus pertinente possible. Plus cette liste est longue, plus il faudra faire d'efforts pour trier les mots de cette liste en cherchant les fonctions biologiques potentielles de tous ces mots. De plus, si un mot w est effectivement sur-représenté pour des raisons biologiques, il entraîne sûrement la sur-représentation d'autres mots de la forme $a\, w_{1:\ell-1}$ ou $w_{2:\ell}\, b$, a, $b \in \mathcal{A}$. On peut alors espérer que ces mots sur-représentés dans le modèle initial aient des fréquences beaucoup mieux prédites dans le modèle conditionné par la fréquence de w.

Pour cela, nous allons construire cette liste de façon itérative. Le premier mot de la liste est le plus mal prédit par le modèle r-markovien. C'est le mot dont l'action associée au PGD est la plus grande. Puis nous conditionnons le modèle r-markovien pour que la fréquence prédite par le modèle corresponde à la fréquence observée. Le deuxième mot de la liste sera le mot dont la fréquence est la plus mal prédite par le modèle conditionné. Pour obtenir un nouveau mot dans la liste, on conditionne donc le modèle pour que les mots précédemment obtenus aient des fréquences correctement prédites par le modèle. Nous cherchons alors le mot dont la fréquence est la plus mal prédite dans ce nouveau modèle.

Nous notons X la chaîne de Markov tensorisée, à valeurs dans $\Sigma = \mathcal{A}^\ell$.

Définition 8.4. *Quelque soit la liste de mots W, on appelle processus de*

comptage des mots de W le processus aléatoire (S_n) défini par

$$S_n = \sum_{j=0}^{n} f(X_j)$$

où f est la fonction dont les coordonnées sont les fonctions indicatrices des éléments de W, $f = (\mathbf{1}_w)_{w\in W}$.

Supposons que W est le début de cette liste, de longueur d. Nous voulons étudier la qualité de la prédiction de la fréquence d'un mot w' de longueur ℓ dans le modèle conditionné. Notons S_n le processus de comptage des mots de W et S'_n le processus de comptage de w'. Alors $T_n = (S_n,\, S'_n)$ est le processus de comptage des mots de $W' = W \cup \{w'\}$. Notons $v = \alpha(W)$ le vecteur des fréquences observées des mots de W et $v' = \alpha(w')$ la fréquence observée du mot w'. On doit étudier

$$p_n := \mathbb{P}\Big(S'_n \approx nv' \,\Big|\, S_n \approx nv\Big),$$

donc

$$\mathbb{P}\Big(T_n \approx nx\Big)$$

où l'on a posé $x = (v,\, v')$. Les corollaires 6.3 et 6.5 nous donnent une estimation de cette probabilité de grandes déviations de variables aléatoires vectorielles. Pour appliquer ce théorème nous devons vérifier l'hypothèse : pour tout vecteur e de la base canonique de $\mathbb{R}^d$, il existe deux mots a et b dans Σ, un vecteur x de $\mathbb{Z}^d$ et un rang n tel que

$$\mathbb{P}_a(T_n = x\,;\, X_n = b) > 0 \quad \text{et}$$
$$\mathbb{P}_a(T_n = x + e\,;\, X_n = b) > 0.$$

Remarquons que cette hypothèse n'est pas vraie quelque soit les ensembles W et le mot w'. En particulier, si la liste W' contient trop de mots, on ne peut pas changer la valeur de T_n sans changer X_0 ou X_n.

Notons $I(W)$ la vitesse de décroissance exponentielle avec n de $\mathbb{P}_a(S_n \approx nv)$ donnée par le PGD de (S_n/n). De même, notons $I(W')$ la vitesse de décroissance exponentielle de $\mathbb{P}_a(T_n \approx nx)$. En appliquant les corollaires 6.3 et 6.5 à S_n et T_n, on obtient un développement de $\log p_n$ comparable à celui obtenu dans le cas de l'étude d'un seul mot dans la proposition 8.3

$$\log p_n = -nI(w'|W) - \frac{1}{2}\log n + O(1).$$

Le terme dominant $I(w'|W)$ est donné par la différence $I(W') - I(W)$. Le terme donné par la différence des produits scalaires

$$-u\,(nx - x_n) + t\,(nv - v_n)$$

est borné par une constante, où $t \in \mathbb{R}^d$ et $u \in \mathbb{R}^{d+1}$ sont des vecteurs de coordonnées strictement positives et les suites (x_n) et (v_n) sont définies, coordonnées par coordonnées comme des parties entières supérieures ou inférieures de nx et nv.

8.3 Mesure empirique des mots de longueur ℓ

Plutôt que de calculer directement $I(w'|W)$ à partir de la valeur propre dominante d'une matrice dépendant d'un paramètre, nous pouvons utiliser le principe de grandes déviations sur la mesure empirique des mots de longueur ℓ, dont la fonction de taux s'exprime plus simplement. On suppose toujours que le modèle aléatoire est r-markovien, de matrice de transition Q et que l'on s'intéresse aux mots de longueur ℓ suffisamment grande, $\ell > r+1$. La condition $\ell > r+1$ revient à dire que l'on a estimé les paramètres du modèle avec les fréquences des mots de longueur strictement inférieure à ℓ.

Définition 8.5. *Notons χ_n la mesure empirique des mots de longueurs ℓ sur le modèle r-markovien, qui est définie pour tout $m \in \Sigma$ avec*

$$\chi_n(m) = \frac{1}{n}\sum_{j=0}^{n-1} \mathbf{1}_m(Y_{j:j+\ell-1}).$$

Avant d'énoncer le principe de grandes déviations auquel χ_n satisfait, donnons quelques définitions préliminaires.

Définition 8.6. *On dit que la mesure $\nu \in \mathcal{M}(\Sigma)$ est invariante par translation si*

$$\forall m \in \mathcal{A}^{\ell-1}, \quad \nu(\mathcal{A}\, m) = \nu(m\, \mathcal{A}).$$

On note $\mathcal{S}_\ell(\mathcal{A})$ l'ensemble des distributions de $\mathcal{M}_1^+(\Sigma)$ invariantes par translation.

Introduisons sur l'ensemble $\mathcal{M}^+(\Sigma)$ des mesures positives sur Σ la fonction J à valeurs dans $[0;+\infty]$ définie par

$$J(\nu) = \sum \nu(m) \log\left(\frac{\nu(m)}{\nu(m_{1:\ell-1}\mathcal{A})Q(m_{\ell-r:\ell-1}, m_\ell)}\right),$$

où l'on somme sur tous les mots m de longueur ℓ avec la convention $0\log 0 = 0$ et $0\log(x/0) = 0$ pour tout $x \in \mathbb{R}$. Nous prolongeons cette fonction sur $\mathcal{M}(\Sigma)$ par $J(\nu) = +\infty$ si $\nu \notin \mathcal{M}^+(\Sigma)$. Munissons l'espace $\mathcal{M}(\Sigma)$ des mesures signées sur Σ de sa topologie d'espace vectoriel normé, en l'identifiant à $\mathbb{R}^\Sigma$. Comme le modèle aléatoire est r-markovien, et $r+1 < \ell$, on obtient le théorème suivant. C'est une conséquence directe du PGD pour la mesure empirique des paires d'états d'une chaîne de Markov finie, voir par exemple [21, section 3.1.3]. En effet, la mesure empirique χ_n est la mesure empirique des paires d'état de la chaîne de Markov tensorisée à valeurs dans $\mathcal{A}^{\ell-1}$.

Théorème 8.7. *La suite des mesures empiriques $(\chi_n)_n$ des mots de longueur ℓ suit un principe de grandes déviations dans $\mathcal{M}(\Sigma)$ de bonne fonction de taux K donnée par*

$$K(\nu) = \begin{cases} J(\nu) & \text{si } \nu \in \mathcal{S}_\ell(\mathcal{A}), \\ +\infty & \text{sinon.} \end{cases} \tag{8.1}$$

Remarque 8.1. Il est facile de comprendre que $K(\nu)$ est infinie en dehors de $\mathcal{S}_\ell(\mathcal{A})$. En effet, pour tout $m \in \mathcal{A}^{\ell-1}$, $\chi_n(m\,\mathcal{A})$ et $\chi_n(\mathcal{A}\,m)$ sont égaux ou diffèrent à $1/n$ près, suivant les valeurs initiales et finales de la chaîne de Markov.

$$\chi_n(m\,\mathcal{A}) - \chi_n(\mathcal{A}\,m) = \frac{1}{n}\Big(\mathbf{1}_{m\,\mathcal{A}}(X_0) - \mathbf{1}_{\mathcal{A}\,m}(X_n)\Big) \tag{8.2}$$

Supposons que $\nu \notin \mathcal{S}_\ell(\mathcal{A})$. Alors, il existe un voisinage B de ν qui ne rencontre pas $\mathcal{S}_\ell(\mathcal{A})$. Si ce voisinage est suffisamment petit, l'équation (8.2) montre qu'il existe un rang n à partir duquel $\chi_n \in B$ n'est jamais réalisable. Donc la probabilité que $\chi_n \in B$ est nulle à partir d'un certain rang. Ce qui montre que $I(\nu) = +\infty$.

Remarque 8.2. Les fonctions J et les fonctions K sont finies sur des ensembles différents. En particulier, la fonction J est finie sur l'ensemble $\mathcal{M}^+(\Sigma)$, d'intérieur non vide dans $\mathcal{M}(\Sigma)$. Et J est de classe C^1 sur cet intérieur. En revanche, la fonction K est finie sur $\mathcal{S}_\ell(\mathcal{A})$ qui est un ensemble d'intérieur vide dans $\mathcal{M}(\Sigma)$.

En appliquant le principe de contraction, on obtient un PGD pour le processus vectoriel (S_n) de comptage de la liste W. Et la fonction de taux $\Lambda^\star$ de ce PGD est définie, pour tout $x = (x_w) \in \mathbb{R}^W$ par

$$\Lambda^\star(x) = \inf\{K(\nu)\,;\, \nu \in \mathcal{M}_1^+(\Sigma) \text{ et } \forall w \in W,\, \nu(w) = x_w\}. \tag{8.3}$$

Remarque 8.3. Nous savons par ailleurs que $\Lambda^\star$ est la transformée de Fenchel-Legendre de la valeur propre dominante de la matrice $Q(t)$. La formule (8.3) permet donc de calculer la valeur propre dominante d'une matrice positive, irréductible et apériodique. On obtient alors la caractérisation de Varadhan du rayon spectral d'une matrice de Perron-Frobenius.

8.4 Programmation

Nous avons réalisé un programme qui fait la chose suivante. On lui donne en entrée l'alphabet, la séquence d'ADN, l'ordre r du modèle markovien et une liste $W = \{w_i\}$ de mots de même longueur ℓ. Le programme renvoie alors la valeur du minimum de l'action sur l'ensemble des ν, mesure de probabilités telles que :

$$\forall i, \nu(w_i) = \text{fréquence observée de } w_i.$$

Même si nous avons optimisé ce programme en temps de calcul, il faut compter entre 3 secondes et 10 secondes pour obtenir la valeur du minimum pour un mot de longueur 5 sur le serveur du laboratoire. La complexité de l'algorithme est exponentielle en la longueur ℓ des mots étudiés.

On identifie l'ensemble des mesures signées $\mathcal{M}(\Sigma)$ à $\mathbb{R}^d$, que l'on munit de sa structure d'espace euclidien.

Définition 8.8. *Nous notons α le vecteur des fréquences observées des mots de longueur ℓ. Soit W une liste de mots. Nous introduisons $E(W)$ l'ensemble des mesures signées, invariantes par translation, de masse totale* 1 *et qui coïncident avec α sur W. C'est un espace affine de $\mathcal{M}(\Sigma)$. On note $\mathbf{T}E(w)$ son espace vectoriel tangent : c'est l'ensemble des mesures signées, invariantes par translation, de masse nulle, et qui sont identiquement nulles sur W.*

La mesure α appartient presque à $E(W)$. En effet, pour tout $m \in \mathcal{A}^{\ell-1}$, $\alpha(m\,\mathcal{A})$ est la fréquence de m dans la séquence observée, à ceci près que l'on ne compte pas la dernière occurrence m, si celle-ci a lieu à la fin de la séquence observée. De même, $\alpha(\mathcal{A}\,m)$ est la fréquence de m dans la séquence observée, à ceci près que l'on ne compte pas la première occurrence de m si celle-ci a lieu au début de la séquence observée. Quand on allonge la séquence observée de ℓ lettres pour que les ℓ premières lettres coïncident avec les ℓ dernières, la mesure α' des fréquences de mots de longueur ℓ est dans $E(W)$. Comme ℓ est négligeable devant la longueur n de la séquence observée, on obtient une distribution α' proche de α. Dans toute la suite, nous remplaçons donc α par α', l'erreur commise étant négligeable.

La vitesse $I(W)$ de décroissante exponentielle de $\mathbb{P}_\mu(S_n \approx n\alpha(W))$, où (S_n) est le processus de comptage de W est le minimum de la fonction K, définie dans le théorème 8.7, sur l'ensemble $E(W)$. La fonction K coïncide avec J sur $E(W)$. De plus, si les coordonnées de μ sont positives, J est C^1 en μ. Le minimum de K sur $E(W)$ est donc atteint au point μ pour lequel le gradient est orthogonal à l'espace tangent $\mathbf{T}E(W)$. Dans ce cas, μ est le vecteur des fréquences des mots de ℓ prédites par le modèle conditionné par les fréquences de W. Notons p la projection orthogonale sur l'espace $\mathbf{T}E(W)$. Nous appliquons l'algorithme de descente du gradient à pas constant en introduisant la suite (ν_k) définie par

$$\nu_{k+1} = \nu_k - \delta\, p\,(J'(\nu_k)),$$

$\delta > 0$. De plus, nous choisissons $\nu_0 = \alpha$ car $\alpha \in E(W)$ quel que soit W.

Reste à calculer la matrice de la projection p. Celle-ci ne dépend pas de l'étape k de l'algorithme de minimisation. Il suffit en fait de savoir projeter sur l'espace vectoriel F orthogonal à $\mathbf{T}E(W)$. Pour cela, on construit une base orthonormale de F dans $\mathbb{R}^d$. La définition de $\mathbf{T}E(W)$ donne directement une famille génératrice de F. Quitte à supprimer quelques vecteurs de cette famille, on obtient une base de F. Pour orthonormaliser cette base, nous utilisons un algorithme de Gram-Schmidt.

Remarque 8.4. En ordonnant correctement les vecteurs de cette base, les $|\Sigma|^{\ell-1}$ premiers vecteurs de cette base forment une base de l'espace orthogonal de $\mathbf{T}E(\emptyset)$, l'espace vectoriel des mesures signées sur Σ, invariantes par translation et de masse nulle. Ils ne dépendent pas de l'ensemble W choisi. Donc, si on veut minimiser l'action K sur $E(W)$, pour différentes listes de mots W, il suffit

de modifier les derniers vecteurs de la base orthonormalisée. Nous enregistrons donc le début de cette base orthonormée pour ne pas avoir à la recalculer à chaque fois que l'on change W.

La partie calcul vectoriel de ce programme est implémentée en C. Pour cela, j'utilise les librairies suivantes :
- GSL (*GNU Scientific Library*), cf `http ://sources.redhat.com/gsl/`. Cette librairie permet de manipuler très simplement des variables de type vecteur et matrice en C.
- BLAS (*Basic Linear Algebra Subprograms*). C'est une librairie de routines qui effectuent les opérations de base impliquant des matrices et des vecteurs. Précisément, j'utilise la version optimisée pour Pentium III (les deux processeurs du serveur du laboratoire sont des Pentium III) fournie par le projet ATLAS.

Le reste constitue un ensemble de fonctions en Python.

8.5 Résultats sur des génomes entiers

J'ai fait tourner ce programme sur la séquence d'ADN complète d'*Escherichia coli* K12. Les listes de mots de longueur 4 obtenues pour le modèle 2-markovien et les listes de mots de longueur 5 pour le modèle 3-markovien sont décevantes. En fait, dans ces deux cas de figure, les signaux que l'on cherche à détecter sont sûrement trop proches du bruit que la séquence aléatoire modélise.

En revanche, nous obtenons des résultats qui semblent pertinents pour les mots de longueur 4 sur le modèle 1-markovien, voir figure 8.1, et pour les mots de longueur 5 sur le modèle 2-markovien, voir figure 8.2. J'ai également fait tourner ce programme sur la séquence d'ADN complète de Bacillus Subtilis, pour obtenir des mots de longueur 5 sur le modèle 2-markovien, voir figure 8.3.

On note $W(k)$ les k premiers mots de la liste de mots exceptionnels, $I(W(k))$ la valeur de l'action associée aux déviations des fréquences des mots de $W(k)$ par rapport au modèle r-markovien et $\nu_{W(k)}$ le vecteur des fréquences des mots de longueur ℓ prédites par le modèle conditionné par $W(k)$.

8.5.1 Les mots de longueur 4 sur le modèle 1-markovien

Les résultats que j'obtiens sur les mots de longueur 4 relativement au modèle 1-markovien sont donnés dans la figure 8.1. Faisons quelques commentaires sur ces résultats.

Dans le modèle 1-markovien, $K(\alpha) = 212 \times 10^{-4}$, où α est la mesure em-

pirique des mots de longueur 4. C'est l'action associée aux déviations des fréquences de tous les mots de longueur 4 dans le modèle 1-markovien. On n'annule donc pas tout phénomène de grandes déviations sur les mots de longueur 4 en conditionnant par les fréquences empiriques de ces trente premiers mots : $I(W(30)) = 143.6 \times 10^{-4}$. Mais une grosse moitié de l'action est expliquée par ces 30 premiers mots, parmi les $4^4 = 256$ mots de longueur 4 sur l'alphabet $\mathcal{A} = \{\texttt{a}, \texttt{c}, \texttt{g}, \texttt{t}\}$.

Rappelons que le motif chi d'E.coli est `gctggtgg`. C'est un mot qui est sur-représenté dans le génome relativement aux modèles statistiques. De plus, on connaît le rôle biologique de ce mot. On retrouve des sous-mots de ce motif, ou de son complémentaire (`ccaccag`) dans cette liste : le premier mot de cette liste (`ctgg`)est un sous-mot du motif chi, et le deuxième mot de cette liste (`ccag`) est son complémentaire. Ils sont tous les deux sur-représentés dans le génome par rapport aux prédictions du modèle M1. Le 25$^{\text{e}}$ mot de cette liste (`ggtg`), qui est un mot sur-représenté, est aussi un sous-mot du motif chi. On trouve son complémentaire `cacc` à la 28$^{\text{e}}$ position dans cette liste.

Notons également que les 7$^{\text{e}}$ et 8$^{\text{e}}$ mots de cette liste sont deux mots sur-représentés qui sont les complémentaires l'un de l'autre : `tcag` et `ctga`. Je ne connais aucune explication à cette sur-représentation. Le même phénomène apparaît également pour les 11$^{\text{e}}$ et 13$^{\text{e}}$ mots de cette liste : `accg` et `cggt`.

Pour les 22 mots sous-représentés de cette liste, on peut faire les remarques suivantes :
- 4 d'entre eux sont leurs propres complémentaires ($w_3 =$ `ctag`, $w_4 =$ `ggcc`, $w_{14} =$ `catg`, $w_{20} =$ `aatt`)).
- 10 d'entre eux vont par paires de mots complémentaires et ces paires sont des mots rapprochés dans la liste (par exemple $w_5 =$ `caag` et $w_6 =$ `cttg` ou $w_{10} =$ `ggac` et $w_{12} =$ `gtcc`). La seule paire où les mots sont éloignés est : $w_9 =$ `ccca` et $w_{22} =$ `tggg`.
- 8 d'entre eux n'ont pas leurs complémentaires dans cette liste : w_{17}, w_{18}, w_{19}, w_{21}, w_{23}, w_{26}, w_{29}, w_{30}.

Parmi les faits remarquables, on peut constater que les paires de mots complémentaires de cette liste sont soit des paires de mots sur-représentés, soit des paires de mots sous-représentés.

En revanche, je ne sais pas quoi penser de $w_3 =$ `ctag`, qui est sous-représenté dans le génome par rapport au modèle initial et sur-représenté lorsqu'on conditionne le modèle par les fréquences empiriques de $w_1 =$ `ctgg` et $w_2 =$ `ccag`.

8.5.2 Les mots de longueur 5 sur le modèle 2-markovien

Les résultats obtenus sur les mots de longueur 5 relativement au modèle 2-markovien du génome d'E.coli sont donnés dans la figure 8.2. Vu ce qui a été constaté sur les mots exceptionnels de longueur 4 par rapport au modèle 1-markovien, on peut commencer par regrouper ces mots par paire de mots complémentaires.

– Il y a quatre paires de mots complémentaires qui sont sur-représentés dans le génome par rapport au modèle 2-markovien : $w_6 = \texttt{ggcga}$ et $w_{10} = \texttt{tcgcc}$, $w_8 = \texttt{ggcaa}$ et $w_{12} = \texttt{ttgcc}$, $w_{11} = \texttt{gaaga}$ et $w_{13} = \texttt{tcttc}$, $w_{23} = \texttt{ggtga}$ et $w_{25} = \texttt{tcacc}$.
– Il y a cinq paires de mots complémentaires qui sont sous-représentés dans le génome par rapport au modèle 2-markovien : $w_1 = \texttt{tccaa}$ et $w_3 = \texttt{ttgga}$, $w_7 = \texttt{cttgg}$ et $w_9 = \texttt{ccaag}$, $w_{15} = \texttt{ggacc}$ et $w_{18} = \texttt{ggtcc}$, $w_{16} = \texttt{ggttc}$ et $w_{17} = \texttt{gaacc}$, $w_{21} = \texttt{gattc}$ et $w_{22} = \texttt{gaatc}$.
– Il n'y a pas de paires de mots complémentaires dont l'un soit sur-représenté et l'autre sous-représenté.

Pour la plupart des mots de cette liste, $\nu_{M2}(w_k)$ est voisin de $\nu_{W(k-1)}(w_k)$. Il y a quelques exceptions que je voudrais passer en revue. Le premier de ces mots est $w_5 = \texttt{gccag}$: la fréquence prédite par le modèle conditionné $\nu_{W(4)}(\texttt{gccag})$ est plus petite que celle prédite par le modèle M2, alors que ce mot est sur-représenté dans le génome. Ce qui veut dire que le conditionnement du modèle par les fréquences de w_1, w_2 et w_3 fait évoluer la fréquence prédite de w_4 dans le mauvais sens par rapport à la fréquence observée. Regardons plus précisément comment évolue cette fréquence prédite à chaque étape du conditionnement : $\nu_{W(1)}(\texttt{gccag}) = 20.72 \times 10^{-4}$, $\nu_{W(2)}(\texttt{gccag}) = 20.776 \times 10^{-4}$, $\nu_{W(3)}(\texttt{gccag}) = 20.878 \times 10^{-4}$ et $\nu_{W(4)}(\texttt{gccag}) = 18.573 \times 10^{-4}$. C'est donc la sous-représentation de $w_4 = \texttt{ggcca}$ qui fait diminuer la fréquence prédite par les modèles conditionnés. On peut remarquer que ces deux mots w_4 et w_5 partagent le sous-mot $\texttt{gcca}$.

Le mot $w_{19} = \texttt{agctg}$ est exactement dans le même cas si ce n'est que sur- et sous-représentation sont inversés : la fréquence prédite par le modèle conditionné est plus grande la fréquence prédite par le modèle 2-markovien initial, alors que ce mot est sous-représenté dans le génome. On peut faire les mêmes calculs pour regarder comment la fréquence prédite évolue quand on change le conditionnement : $\nu_{W(i)}(\texttt{agctg})$ pour $1 \leq i \leq 13$ sont tous compris entre 15.94×10^{-4} et 16.33×10^{-4}. Ensuite il y a un saut entre $\nu_{W(13)}(\texttt{agctg}) = 15.947 \times 10^{-4}$ et $\nu_{W(14)}(\texttt{agctg}) = 17.551 \times 10^{-4}$. Et $\nu_{W(i)}(\texttt{agctg})$ augmente légèrement avec i entre $i = 14$ et $i = 18$ pour atteindre finalement $\nu_{W(18)}(\texttt{agctg}) = 17.71 \times 10^{-4}$. On peut en conclure que c'est la sur-représentation de $w_{14} = \texttt{gctgg}$ qui augmente nettement la fréquence dans les modèles conditionnés. Et on remarque que ce mot partage avec $w_{19} = \texttt{agctg}$ le sous-mot $\texttt{gctg}$.

k	**(a)**	**(b)** $\times 10^{-4}$	**(c)** $\times 10^{-4}$	**(d)**	**(e)** $\times 10^{-3}$	**(f)** $\times 10^{-3}$	**(g)** $\times 10^{-3}$
1	ctgg	17.9875	17.9875	+	3.297		7.272
2	ccag	17.3056	35.2931	+	3.361	3.402	7.385
3	ctag	12.7508	48.0440	?	1.968	1.888	1.911
4	ggcc	7.5169	55.5610	−	4.373	5.179	2.709
5	caag	6.5521	62.1131	−	4.316	4.167	2.089
6	cttg	6.7761	68.8893	−	4.275	4.187	2.073
7	tcag	5.6146	74.5040	+	3.307	3.257	5.310
8	ctga	5.5015	80.0055	+	3.262	3.223	5.251
9	ccca	5.1850	85.1906	−	3.717	5.267	2.931
10	ggac	4.2990	89.4896	−	2.970	3.282	1.777
11	accg	4.6892	94.1788	+	3.744	3.536	5.448
12	gtcc	3.8304	98.0093	−	2.972	3.184	1.782
13	cggt	3.6126	101.6220	+	3.724	3.623	5.317
14	catg	3.1409	104.7629	−	5.368	4.870	3.286
15	ttag	2.7701	107.5331	−	2.830	2.739	1.609
16	ctaa	2.7749	110.3080	−	2.795	2.768	1.629
17	cgag	2.6782	112.9863	−	3.533	3.708	2.286
18	gcac	2.5382	115.5245	−	5.125	5.191	3.125
19	gctc	2.3778	117.9024	−	3.879	4.632	2.675
20	aatt	2.3587	120.2611	−	5.877	5.896	4.358
21	gtgc	2.4032	122.6644	−	5.075	5.191	3.674
22	tggg	2.3371	125.0016	−	3.658	4.734	2.886
23	cctc	2.1784	127.1800	−	2.745	2.529	1.377
24	gtgt	2.2384	129.4185	−	3.379	3.766	2.372
25	ggtg	2.6783	132.0968	+	3.571	3.665	5.129
26	tgag	2.2212	134.3181	−	3.284	3.664	2.518
27	acac	2.2722	136.5904	−	3.427	3.598	2.363
28	cacc	3.1416	139.7320	+	3.627	3.626	5.058
29	ggag	2.0236	141.7557	−	2.753	3.104	2.684
30	gcat	1.8675	143.6233	−	6.866	6.199	4.736

FIG. 8.1 – **Liste des 30 premiers mots exceptionnels de longueur 4 sur le modèle 1-markovien du génome d'E.Coli.**
On note $W(k) = \{w_i\,;\, 1 \leq i \leq k\}$ et $I(W(k))$ la valeur de l'action qui nous permet de quantifier la qualité de la prédiction des fréquences des mots de $W(k)$ dans le modèle 1-markovien.

- **(a)** w_k est le dernier mot exceptionnel de la liste à l'étape k ;
- **(b)** $I(W(k)) - I(W(k-1))$ est l'action dans le modèle conditionné ;
- **(c)** $I(W(k))$, action dans le modèle initial ;
- **(d)** + si le mot est sur-représenté dans le génome par rapport aux modèles, − s'il est sous-représenté, ? s'il change de statut entre le modèle initial et le modèle conditionné ;
- **(e)** $\nu_{M1}(w_k)$ fréquence moyenne d'apparition de w_k dans le modèle initial ;
- **(f)** $\nu_{W(k-1)}(w_k)$ fréquence moyenne d'apparition de w_k dans le modèle conditionné par les fréquences d'apparition des mots de $W(k-1)$;
- **(g)** $\alpha(w_k)$ fréquence d'apparition de w_k dans le génome.

k	**(a)**	**(b)** $\times 10^{-4}$	**(c)** $\times 10^{-4}$	**(d)**	**(e)** $\times 10^{-4}$	**(f)** $\times 10^{-4}$	**(g)** $\times 10^{-4}$
1	tccaa	3.031	3.031	−	9.054		2.763
2	ggccg	2.975	6.006	−	15.406	15.506	6.953
3	ttgga	3.002	9.008	−	9.121	9.246	2.892
4	ggcca	2.107	11.116	−	15.302	15.585	8.227
5	gccag	2.185	13.302	+	20.551	18.573	27.97
6	ggcga	1.943	15.245	+	12.132	12.373	19.839
7	cttgg	1.917	17.163	−	8.219	6.987	2.524
8	ggcaa	1.939	19.102	+	11.702	11.769	19.065
9	ccaag	1.741	20.843	−	8.234	6.758	2.56
10	tcgcc	1.778	22.622	+	12.495	13.221	20.514
11	gaaga	1.697	24.319	+	8.034	8.078	13.935
12	ttgcc	1.707	26.027	+	11.873	12.363	19.287
13	tcttc	1.732	27.759	+	7.964	7.92	13.786
14	gctgg	1.528	29.288	+	19.934	19.827	27.909
15	ggacc	1.281	30.569	−	7.24	6.766	3.058
16	ggttc	1.283	31.853	−	12.78	12.761	7.516
17	gaacc	1.231	33.085	−	12.837	12.728	7.578
18	ggtcc	1.197	34.283	−	7.121	6.726	3.136
19	agctg	1.182	35.465	−	15.89	17.711	11.9
20	ccgga	1.148	36.614	+	9.782	9.062	13.926
21	gattc	1.168	37.782	−	12.404	12.746	7.74
22	gaatc	1.175	38.958	−	12.34	12.749	7.727
23	ggtga	1.171	40.129	+	10.731	10.664	15.97
24	tccag	1.137	41.267	+	12.385	11.709	17.151
25	tcacc	1.098	42.365	+	10.853	10.659	15.772
26	ggccc	1.027	43.393	−	8.455	8.407	4.617
27	caaca	0.974	44.367	+	9.22	9.122	13.666
28	tgatg	0.962	45.329	+	14.34	14.51	20.149
29	ctgag	0.928	46.258	−	9.131	8.607	4.955
30	ctaga	0.901	47.159	−	1.764	1.707	0.308
31	tcgca	0.898	48.057	−	12.905	12.523	8.195

FIG. 8.2 – **Liste des 31 premiers mots exceptionnels de longueur 5 sur un modèle 2-markovien du génome d'E.coli.**
On note $W(k) = \{w_i\,;\, 1 \leq i \leq k\}$ et $I(W(k))$ la valeur de l'action qui nous permet de quantifier la qualité de la prédiction des fréquences des mots de $W(k)$ dans le modèle 2-markovien.

- **(a)** w_k, dernier mot exceptionnel de la liste à l'étape k ;
- **(b)** $I(W(k)) - I(W(k-1))$, action dans le modèle conditionné ;
- **(c)** $I(\nu_{W(k)})$, action dans le modèle initial ;
- **(d)** + si le mot est sur-représenté dans le génome par rapport aux modèles, − s'il est sous-représenté ;
- **(e)** $\nu_{M2}(w_k)$ fréquence moyenne d'apparition de w_k dans le modèle initial ;
- **(f)** $\nu_{W(k-1)}(w_k)$ fréquence moyenne d'apparition de w_k dans le modèle conditionné par les fréquences d'apparition des précédents mots de la liste ;
- **(g)** $\alpha(w_k)$ fréquence d'apparition de w_k dans le génome.

k	**(a)**	**(b)** $\times 10^{-5}$	**(c)** $\times 10^{-5}$	**(d)**	**(e)** $\times 10^{-4}$	**(f)** $\times 10^{-4}$	**(g)** $\times 10^{-4}$
1	aattt	15.1673	15.1673	−	27.4394		19.2283
2	aatgg	11.601	26.768	+	10.1458	10.2623	15.2849
3	ccatt	11.619	38.387	+	10.1683	10.306	15.273
4	tcaga	9.644	48.031	−	16.6137	16.2116	11.3675
5	tctga	8.757	56.788	−	16.8457	16.4414	11.8777
6	cttct	8.301	65.09	+	12.8021	12.9872	17.2993
7	cggca	7.731	72.821	+	11.1174	11.2076	15.3726
8	cgatt	7.843	80.664	+	10.4402	10.5072	14.5351
9	aactt	7.582	88.246	−	12.1182	12.1541	8.3447
10	cagca	7.172	95.418	+	14.9166	15.43	19.736
11	tgctg	7.169	102.587	+	14.8231	15.3492	19.5272
12	gaaga	6.675	109.261	+	14.5802	15.0252	18.8486
13	aagat	9.118	118.379	−	18.389	18.8203	14.2077
14	atctt	6.28	124.659	−	18.6048	18.3875	14.2148
15	aaatt	6.554	131.213	−	27.3024	23.2402	19.1405
16	tcttc	6.502	137.715	+	14.9256	16.1612	19.7669
17	ccagg	6.001	143.716	−	6.3774	6.2352	3.7868
18	cctgg	6.459	150.175	−	6.5361	6.4091	3.9624
19	aaaga	5.616	155.791	+	23.5117	22.7034	26.1945
20	tgccg	5.5	161.291	+	11.0514	11.3799	14.8198
21	gatga	5.386	166.677	+	15.4326	15.5514	19.0954
22	aatcg	5.98	172.657	+	10.6185	11.0892	14.2883
23	aggag	5.162	177.819	+	6.5758	6.509	9.1443
24	tcttt	5.524	183.343	+	23.969	23.4843	26.8517
25	aagtt	4.77	188.114	−	11.9245	11.5975	8.6105
26	tcatc	5.389	193.503	+	15.7591	15.9618	19.0764
27	ggcgg	4.389	197.892	+	7.8995	8.1615	10.8171
28	aaacg	4.403	202.295	+	14.0469	14.9484	17.9185
29	ctcct	5.265	207.56	+	6.7993	6.7856	9.1894
30	ccgcc	4.546	212.106	+	7.8738	8.1182	10.7483

FIG. 8.3 – **Liste des 30 premiers mots exceptionnels de longueur 5 sur un modèle 2-markovien du génome de Bacillus Subtilis.**
On note $W(k) = \{w_i\,;\, 1 \leq i \leq k\}$ et $I(W(k))$ la valeur de l'action qui nous permet de quantifier la qualité de la prédiction des fréquences des mots de $W(k)$ dans le modèle 2-markovien.

- **(a)** w_k, dernier mot exceptionnel de la liste à l'étape k ;
- **(b)** $I(W(k)) - I(W(k-1))$, action dans le modèle conditionné ;
- **(c)** $I(\nu_{W(k)})$, action dans le modèle initial ;
- **(d)** + si le mot est sur-représenté dans le génome par rapport aux modèles, − s'il est sous-représenté ;
- **(e)** $\nu_{M2}(w_k)$ fréquence moyenne d'apparition de w_k dans le modèle initial ;
- **(f)** $\nu_{W(k-1)}(w_k)$ fréquence moyenne d'apparition de w_k dans le modèle conditionné par les fréquences d'apparition des précédents mots de la liste ;
- **(g)** $\alpha(w_k)$ fréquence d'apparition de w_k dans le génome.

Questions ouvertes

Généralisation aux espaces d'états infinis

Les estimations obtenues dans les chapitres 3 à 6 sont limitées aux chaînes de Markov définies sur un espace d'états fini. La généralisation de ces résultats aux chaînes de Markov sur un espace d'états dénombrable est une question naturelle. Même pour un espace d'états dénombrable, la généralisation ne semble pas évidente. En effet, Bryc et Smoleński [15] ont donné un exemple de chaîne de Doeblin irréductible et récurrente sur un espace d'états dénombrable, mais qui ne satisfait pas à un principe de grandes déviations (PGD).

Iscoe et al. [31] et Ney et Nummelin. [45] montrent que la technique de construction de la chaîne de Markov twistée s'adapte aux chaînes de Harris récurrentes, pour lesquelles on peut définir Λ sur un ouvert contenant l'origine de $\mathbb{R}^d$. Kontoyiannis et Meyn [35] ont généralisé certains résultats aux chaînes de Doeblin récurrentes $(X_n)_{n\in\mathbb{N}}$ et obtiennent des estimations précises de grandes déviations sur $S_n = \sum_{k=0}^{n-1} f(X_k)$ lorsque f est à valeurs réelles. Dans le cas où l'espace d'états est infini, il faut faire attention à l'espace des fonctions de Σ dans $\mathbb{R}$ ou $\mathbb{C}$ sur lequel on étudie les propriétés spectrales des noyaux $Q(t)$. Ils supposent que (X_n) est géométriquement ergodique, de fonction de Lyapunov V, et ils utilisent l'espace de fonctions L_V^∞ pour faire l'étude spectrale de $Q(t)$, qui est défini par

$$L_V^\infty := \{f : \Sigma \to \mathbb{C}\,;\, \sup |f(z)/V(z)| < \infty\}\,.$$

Ils obtiennent ainsi des estimations des probabilités de $\{S_n \geq nx\}$ pour des valeurs de x suffisamment proche de la dérive moyenne $\lim n^{-1}S_n$. Mais il n'y a aucune chance de pouvoir étendre ces résultats à toutes les valeurs de x comme dans le cas où l'espace d'états est fini, puisque Bryc et Smoleński ont donné un contre-exemple.

Généralisation pour d'autres formes de B

Dans le chapitre 6, nous avons donné des résultats asymptotiques pour $\mathbb{P}_a(S_n \in n\,B)$ quand B est un produit de demi-droite. Dans ce cas, l'ensemble B est anguleux au point dominant. Avec les résultats de cette thèse, il me semble possible d'obtenir des résultats sur les processus de Markov additifs qui généralisent ceux de Ney [44] et Iltis [30] sur les sommes de variables i.i.d. En particulier, lorsque la frontière de B est lisse autour du point dominant v, la probabilité $\mathbb{P}_a(S_n \in n\,B)$ est-elle équivalente à

$$C\,n^{-\gamma}\mathrm{e}^{-n\Lambda^\star(v)}\,?$$

Le cas échéant, quelles sont les valeurs de γ et de C ?

En revanche, les résultats de Barbe et Broniatowski [6] sont plus délicats à adapter à notre contexte. Leurs résultats reposent sur un lemme qui construit la distribution twistée pour tous les points de l'ensemble B. Précisément, si *(i)* (X_k) est une suite de variables aléatoires i.i.d. à valeurs dans $\mathbb{R}^d$, si *(ii)* $S_n = \sum_{k=0}^{n-1} X_k$, si *(iii)* Λ est la fonction génératrice des cumulants de X_0, si *(iv)* Ξ est l'homéomorphisme inverse de Λ', et si *(v)* $\Lambda^\star$ la transformée de Fenchel-Legendre de Λ, alors le lemme de Barbe et Broniatowski montre que, pour tout borélien $B \subset R^d$,

$$\int\int \mathrm{e}^{x\,\Xi((s-x)/n)}\mathbf{1}_B\left(\frac{s-x}{n}\right)\mathrm{d}\mathbb{P}(S_n = s)\mathrm{d}x = n^d\int_B \mathrm{e}^{-n\Lambda^\star(y)}\mathrm{d}y.$$

Malheureusement, il n'y a aucune chance pour que ce lemme soit vrai dans le cas des processus de Markov additifs, car

$$\mathbb{E}_a(\exp t\,S_n) \neq \exp n\Lambda(t).$$

À partir de quel rang est-on dans le régime asymptotique des grandes déviations ?

Les techniques que nous utilisons dans cette thèse permettent de répondre à cette question, si l'on sait obtenir *(i)* une borne explicite dans le développement de Edgeworth obtenu au chapitre 5 et *(ii)* le rayon de convergence des développement en série entière à l'origine de Λ et G_a.

Pour obtenir les bornes C_k dans le théorème 5.4, nous avons majoré trois intégrales $I_1(n)$, $I_2(n)$ et $I_3(n)$. Ce sont les majorations de $I_2(n)$ et $I_3(n)$ qui sont les moins explicites. En particulier, nous utilisons le corollaire 4.6 de la proposition 4.5 qui donne le comportement de la fonction caractéristique du processus de Markov additif à valeurs dans $\mathbb{Z}^d$, en dehors d'un voisinage de

l'origine. Dans ce corollaire, la constante A n'est pas explicite. Dans l'article de Kontoyiannis et Meyn [35], l'écart entre la densite de $n^{-1/2}(S_n - nv)$ et son approximation n'est malheureusement pas plus explicite : les constantes B_0 dans leurs théorèmes 4.1 et 4.16 ne sont pas données en fonction des paramètres du processus.

De plus, je ne sais pas calculer ou minorer les rayons de convergence des séries entières de $\Lambda(t)$, la valeur propre dominante de $Q(t)$, ainsi que celui des coordonnées du vecteur propre associé $G(t)_a$. Ce rayon de convergence intervient lorsque nous utilisons le lemme 5.5 dans la majoration de $I_1(n)$. Obtenir une minoration de ces rayons de convergence permettra d'obtenir une majoration plus explicite de $I_1(n)$.

Minimisation avec pénalisation pour calculer une valeur approchée de l'action

Dans le chapitre 8, nous devons résoudre numériquement un problème de minimisation convexe avec contrainte pour donner une valeur approchée de l'action. Nous avons proposé une méthode pour calculer cette valeur approchée par un algorithme de descente du gradient. Dans cet algorithme, nous projetons le gradient sur un sous-espace vectoriel afin de respecter les contraintes. Le calcul de la matrice de projection et des projections de chacuns des gradients qui interviennent dans l'algorithme ralentit considérable l'algorithme et augmente l'espace mémoire. C'est pourquoi, en pratique, nous avons des résultats numériques uniquement sur les mots de longueur $\ell \leq 5$ avec des modèles m-markoviens.

Pour éviter cette étape de projection, il faudrait plutôt utiliser un algorithme de minimisation avec pénalisation. Mais, je n'ai pas réussi à régler la pénalisation : soit elle prend le dessus trop vite, et l'algorithme ne voit plus la fonction à minimiser, soit la pénalisation est trop faible et les contraintes ne sont plus respectées.

Bibliographie

[1] Alfred V. Aho and Jeffrey D. Ullman. *Concepts fondamentaux de l'informatique*. Dunod, 1984.

[2] Cristina Andriani and Paolo Baldi. Sharp estimates of deviations of the sample mean in many dimensions. *Ann. Inst. H. Poincaré Probab. Statist.*, 33(3) :371–385, 1997.

[3] Raghu Raj Bahadur and R. Ranga Rao. On deviations of the sample mean. *Ann. Math. Statist.*, 31 :1015–1027, 1960.

[4] August A. Balkema, Claudia Klüppelberg, and Sidney I. Resnick. Densities with Gaussian tails. *Proc. London Math. Soc. (3)*, 66(3) :568–588, 1993.

[5] August A. Balkema, Claudia Klüppelberg, and Sidney I. Resnick. Limit laws for exponential families. *Bernoulli*, 5(6) :951–968, 1999.

[6] Philippe Barbe and Michel Broniatowski. Large-deviation probability and the local dimension of sets. In *Proceedings of the 19th Seminar on Stability Problems for Stochastic Models, Part I (Vologda, 1998)*, volume 99, pages 1225–1233, 2000.

[7] Ole E. Barndorff-Nielsen and Claudia Klüppelberg. Tail exactness of multivariate saddlepoint approximations. *Scand. J. Statist.*, 26(2) :253–264, 1999.

[8] Rabi N. Bhattacharya and R. Ranga Rao. *Normal approximation and asymptotic expansions*. Robert E. Krieger Publishing Co. Inc., Melbourne, FL, 1986. Reprint of the 1976 original.

[9] Harald Bohman. What is the reason that Esscher's method of approximation is as good as it is ? *Skand. Aktuarietidskr.*, 1963 :87–94 (1964), 1964.

[10] Alexander A. Borovkov. The Cramér transform, large deviations in boundary problems, and the conditional invariance principle. *Sibirsk. Mat. Zh.*, 36(3) :493–509, i, 1995.

[11] Alexander. A. Borovkov and Anatolii A. Mogul′skiĭ. Large deviations and testing statistical hypotheses. I. Large deviations of sums of random vectors. *Siberian Adv. Math.*, 2(3) :52–120, 1992. Siberian Advances in Mathematics.

[12] Alexander A. Borovkov and Anatolii A. Mogul′skiĭ. Large deviations and testing statistical hypotheses. II. Large deviations of maximum points of random fields. *Siberian Adv. Math.*, 2(4) :43–72, 1992. Siberian Advances in Mathematics.

[13] Alexander A. Borovkov and Anatolii A. Mogul′skiĭ. Large deviations and testing statistical hypotheses. III. Asymptotically optimal tests for composite hypotheses. *Siberian Adv. Math.*, 3(1) :19–86, 1993. Siberian Advances in Mathematics.

[14] Alexander A. Borovkov and Anatolii A. Mogul′skiĭ. Large deviations and testing statistical hypotheses. IV. The statistical invariance principle and the laws of conservation. *Siberian Adv. Math.*, 3(2) :14–80, 1993. Siberian Advances in Mathematics.

[15] Włodzimierz. Bryc and Włodzimierz Smoleński. On the convergence of averages of mixing sequences. *J. Theoret. Probab.*, 6(3) :473–483, 1993.

[16] Henri Cartan. *Elementary theory of analytic functions of one or several complex variables*. Dover, New-York, 1995. Translated from the French, Reprint of the 1973 edition.

[17] Narasinga R. Chaganty and Jayaram Sethuraman. Strong large deviation and local limit theorems. *Ann. Probab.*, 21(3) :1671–1690, 1993.

[18] Narasinga R. Chaganty and Jayaram Sethuraman. Multidimensional strong large deviation theorems. *J. Statist. Plann. Inference*, 55(3) :265–280, 1996.

[19] Henry E. Daniels. Saddlepoint approximations in statistics. *Ann. Math. Statist.*, 25 :631–650, 1954.

[20] Richard A. Davis and Sidney I. Resnick. Extremes of moving averages of random variables with finite endpoint. *Ann. Probab.*, 19(1) :312–328, 1991.

[21] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications, 2nd ed.*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1998.

[22] Alain Denise, Mireille Régnier, and Mathias Vandenbogaert. Assessing the statistical significance of overrepresented oligonucleotides. In *Algorithms in bioinformatics (Århus, 2001)*, volume 2149 of *Lecture Notes in Comput. Sci.*, pages 85–97. Springer, Berlin, 2001.

[23] Fredrik Esscher. On approximate computation of distribution functions when the corresponding characteristic functions are known. *Skand. Aktuarietidskr.*, 1963 :78–86 (1964), 1964.

[24] William. Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York, 1966.

[25] Philippe Flajolet and Andrew Odlyzko. Singularity analysis of generating functions. *SIAM J. Discrete Math.*, 3(2) :216–240, 1990.

[26] Philippe Flajolet and Michèle Soria. General combinatorial schemes : Gaussian limiting distribution and exponential tails. *Discrete Mathematics*, 114(1-3) :159–180, 1993.

[27] Friedrich Götze and Christian Hipp. Asymptotic expansions for sums of weakly dependent random vectors. *Z. Wahrsch. Verw. Gebiete*, 64(2) :211–239, 1983.

[28] Bronius Grigelionis. Conditionally exponential families and Lundberg exponents of Markov additive processes. In *Probability theory and mathematical statistics (Vilnius, 1993)*, pages 337–350. TEV, Vilnius, 1994.

[29] Marjorie G. Hahn and Michael J. Klass. Approximation of partial sums of arbitrary i.i.d. random variables and the precision of the usual exponential upper bound. *Ann. Probab.*, 25(3) :1451–1470, 1997.

[30] Michael Iltis. Sharp asymptotics of large deviations in $\mathbf{R}^d$. *J. Theoret. Probab.*, 8(3) :501–522, 1995.

[31] Ian Iscoe, Peter Ney, and Esa Nummelin. Large deviations of uniformly recurrent Markov additive processes. *Adv. in Appl. Math.*, 6(4) :373–412, 1985.

[32] Jens L. Jensen. Uniform saddlepoint approximations. *Adv. in Appl. Probab.*, 20(3) :622–634, 1988.

[33] Jens L. Jensen. Uniform saddlepoint approximations and log-concave densities. *J. Roy. Statist. Soc. Ser. B*, 53(1) :157–172, 1991.

[34] Claudia Klüppelberg and Thomas Mikosch. Large deviations of heavy-tailed random sums with applications in insurance and finance. *J. Appl. Probab.*, 34(2) :293–308, 1997.

[35] Ioannis Kontoyiannis and Sean P. Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.*, 13(1) :304–362, 2003.

[36] Anders Krogh, I. Saira Mian, and David Haussler. A Hidden Markov model that finds genes in E.coli DNA. *Nucleic Acids Research*, 22 :4768–4778, 1994.

[37] Soumendra Nath Lahiri. Refinements in asymptotic expansions for sums of weakly dependent random vectors. *Ann. Probab.*, 21(2) :791–799, 1993.

[38] Carlos A. León and François Perron. Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.*, 14(2) :958–970, 2004.

[39] Mikhail A. Lifshits. On the lower tail probabilities of some random series. *Ann. Probab.*, 25(1) :424–442, 1997.

[40] Brad Mann. *Berry-Esseen Central Limit Theorems For Markov Chains.* PhD thesis, Harvard University, 1996.

[41] Makoto Matsumoto and Takuji Nishimura. Mersenne twister : A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1) :3–30, 1998.

[42] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[43] Sergei V. Nagaev. Large deviations of sums of independent random variables. *Ann. Probab.*, 7(5) :745–789, 1979.

[44] Peter Ney. Dominating points and the asymptotics of large deviations for random walk on $\mathbf{R}^d$. *Ann. Probab.*, 11(1) :158–167, 1983.

[45] Peter Ney and Esa Nummelin. Markov additive processes. I. Eigenvalue properties and limit theorems. *Ann. Probab.*, 15(2) :561–592, 1987.

[46] Peter Ney and Esa Nummelin. Markov additive processes II. Large deviations. *Ann. Probab.*, 15(2) :593–609, 1987.

[47] Pierre Nicodème, Bruno Salvy, and Philippe Flajolet. Motif statistics. In *Algorithms—ESA '99 (Prague)*, volume 1643 of *Lecture Notes in Comput. Sci.*, pages 194–211. Springer, Berlin, 1999.

[48] James R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[49] Grégory Nuel. *Grandes déviations et chaînes de Markov pour l'étude des occurrences de mots dans les séquences biologiques*. PhD thesis, Université d'Évry, Val d'Essonne, 2001.

[50] Bernard Prum, François Rodolphe, and Élisabeth de Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *J. Roy. Statist. Soc. Ser. B*, 57(1) :205–220, 1995.

[51] R Development Core Team. *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2003. ISBN 3-900051-00-3.

[52] Mireille Régnier. A unified approach to word occurrences probabilities. *Discrete Applied Mathematics*, 104(1) :259–280, 2000.

[53] Mireille Régnier and Alain Denise. Rare events and conditional events on random strings. *Discrete Math. Theor. Comput. Sci.*, 6(2) :191–213 (electronic), 2004.

[54] Mireille Régnier and Wojciech Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22(4) :631–649, 1998.

[55] Gesine Reinert, Sophie Schbath, and Michael S. Waterman. Probabilistic and Statistical Properties of Words : An Overview. *Journal of Computational Biology*, 7(1-2) :1–46, 2000.

[56] Yosef Rinott and Vladimir Rotar. A multivariate CLT for local dependence with $n^{-1/2} \log n$ rate and applications to multivariate graph related statistics. *J. Multivariate Anal.*, 56(2) :333–350, 1996.

[57] Yosef Rinott and Vladimir Rotar. On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted U-statistics. *Ann. Appl. Probab.*, 7(4) :1080–1105, 1997.

[58] Christine Ritzmann. Good local bounds for simple random walks. Available electronically on math arXiv, 2002.

[59] Stéphane Robin and Jean-Jacques Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probab.*, 36(1) :179–193, 1999.

[60] Sophie Schbath. Compound Poisson approximation of word counts in DNA sequences. *ESAIM Probab. Statist.*, 1 :1–16, 1995/97.

[61] Terry Speed and Jaimyoung Kwon. Simple Statistics of DNA Sequences. *Semaine 11 du cours Stat 260 : Statistics in Genetics*, 1998. Disponible à l'adresse : www.stat.berkeley.edu/users/terry.

[62] Nelly Torrent. *Applications des grandes déviations et de la loi d'Erdös-Rényi pour les variables indépendantes ou de dépendance markovienne.* PhD thesis, Université Paris 7, 1998.

[63] Richard L. Tweedie. R-theory for Markov chains on a general state space. I. Solidarity properties and R-recurrent chains. *Ann. Probab.*, 2 :840–864, 1974.

[64] Richard L. Tweedie. R-theory for Markov chains on a general state space. II. r-subinvariant measures for r-transient chains. *Ann. Probab.*, 2 :865–878, 1974.

[65] Michael S. Waterman. *Introduction to computational biology.* Interdisciplinary Statistics. Chapman and Hall, 1995.

Abstract

To establish lists of words with unexpected frequencies in random sequences, for instance in a molecular biology context, one needs to quantify the exceptionality of families of word frequencies. We study large deviation probabilities of d dimensional word counts S_n in Markov models and hidden Markov models. When $s_n = nv + o(n^{1/2})$ deviates from the expected typical behaviour of S_n, we prove that

$$\mathbb{P}(S_n = s_n) \sim c\, n^{-d/2}\, u_n\, \mathrm{e}^{-n\,\lambda},$$

where c and λ are explicit, positive functions of v. The explicit bounded sequence $\{u_n\}$ describes $(s_n - nv)$ and may lead to some periodic oscillations. Analogous equivalents hold for the exceedance probabilities $\mathbb{P}(S_n \geq s_n)$. To prove these results, we establish Edgeworth-like expansions, namely the fact that, for every nonnegative k,

$$\left| n^{d/2}\, \mathbb{P}(S_n = x) - \sum_{j=0}^{k} n^{-j/2}\, \psi_j\left(\frac{x - nm}{\sqrt{n}}\right) \right| \leq C_k\, n^{-(k+1)/2},$$

where C_k is finite, the function ψ_0 is a d dimensional explicit Gaussian density, and ψ_j are built upon partial derivatives of ψ_0. Finally we provide detailed simulations, which exhibit in particular the periodic oscillations mentioned above and lists of words with unexpected frequencies in the genomic sequences of Escherichia coli and Bacillus subtilis.

Keywords: Markov process, large deviations, Edgeworth expansions, Protein and DNA sequences.

Résumé

Pour obtenir des listes de mots de fréquences exceptionnelles par rapport à un modèle aléatoire, par exemple dans un contexte de biologie moléculaire, il faut quantifier la qualité de la prédiction des fréquences d'une famille de mots. Nous étudions les probabilités de grandes déviations du processus vectoriel S_n de comptage d'une famille de d mots dans des modèles de Markov et des modèles de Markov cachés. Quand $s_n = nv + o(n^{1/2})$ dévie du comportement typique de S_n, nous montrons que

$$\mathbb{P}(S_n = s_n) \sim c\, n^{-d/2}\, u_n\, \mathrm{e}^{-n\,\lambda},$$

où c et λ sont des fonctions positives, explicites qui dépendent de v. La suite (u_n) est donnée par $(s_n - nv)$ et peut conduire à des phénomènes d'oscillations périodiques. Nous obtenons des équivalents semblables pour la probabilité $\mathbb{P}(S_n \geq s_n)$. Pour démontrer ces résultats, nous établissons un développement de Edgeworth, c'est-à-dire, pour tout entier $k \geq 0$,

$$\left| n^{d/2}\, \mathbb{P}(S_n = x) - \sum_{j=0}^{k} n^{-j/2}\, \psi_j\left(\frac{x - nm}{\sqrt{n}}\right) \right| \leq C_k\, n^{-(k+1)/2},$$

où C_k est fini, la fonction ψ_0 est une densité gaussienne d-dimensionnelle et les ψ_j sont construites à partir des dérivées partielles de ψ_0. Ensuite, nous donnons des simulations qui montrent en particulier les oscillations périodiques mentionnées ci-dessus ainsi que des listes de mots de fréquences exceptionnelles sur les génomes d'Escherichia coli et Bacillus subtilis.

Mots-clés: Processus de Markov, grandes déviations, développement de Edgeworth, séquences d'ADN ou de protéine.