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

Quelques applications de la théorie des grandes déviations.

Thèse soutenue publiquement le 6 Décembre 2005

en vue de l'obtention du

Diplôme de Doctorat

(arrêté du 30 mars 1992)

Spécialité : Mathématiques

par

Clément DOMBRY

Composition du jury

Rapporteurs :	**Craig John BENHAM**
	Francis COMETS
Examinateurs :	**Stéphane ATTAL**
	Nadine GUILLOTIN-PLANTARD
	Christian MAZZA

Numéro d'ordre 245 - 2005
Institut Camille Jordan UMR 5208

Mis en page avec la classe thloria.

Résumé de la thèse

Quelques applications de la théorie des grandes déviations.

Cette thèse propose plusieurs applications de la théorie des grandes déviations.

Nous étudions d'abord un modèle de dénaturation de l'ADN proposé par Benham. C'est un modèle de spin 1-dimensionnel en champ extérieur hétérogène. Sous certaines hypothèses sur le champ extérieur, nous prouvons un principe de grandes déviations pour la dénaturation, et en déduisons une loi des grands nombres. Nous explicitons la valeur limite de la dénaturation en fonction de certains paramètres (température, superhélicité etc.).

Dans une seconde partie, nous nous intéressons à un algorithme génétique de sélection-mutation en population infinie sur $\mathbb{Z}$. Nous rattachons ce problème à un modèle de marche pondérée construit à partir de la marche simple sur $\mathbb{Z}$ en affectant chaque trajectoire d'un poids donné par une fonctionnelle multiplicative. Nous prouvons un principe de grandes déviations fonctionnel pour la marche pondérée et en déduisons le comportement asymptotique de l'algorithme génétique.

Enfin, nous introduisons une nouvelle modélisation des structures de données, généralisant le modèle markovien étudié par Flajolet, Louchard ... Afin de permettre une inhomogénéité temporelle dans le modèle markovien, nous introduisons les marches aléatoires dynamiques définies par Guillotin et définissons le modèle dynamique de structures de données. Nous démontrons un principe de grandes déviations fonctionnel pour le processus de taille des structures de données dynamiques et en déduisons une loi des grands nombres fonctionnelle.

Mots-clés : Grandes déviations ; Méthode de Laplace ; Dénaturation de l'ADN ; Algorithme génétique ; Structure de données ; Marche aléatoire dynamique.

Some applications of large deviations theory.

In this dissertation, we present some applications of large deviations theory.

A stochastic model for DNA denaturation is studied in a first part. This is a 1-dimensional spin model in inhomogeneous field proposed by Benham. Under suitable conditions on the external field, we prove a large deviations principle for denaturation and also deduce a law of large numbers. The limit denaturation is explicited as a function of several parameters (temperature, superhelicity etc.).

In a second part, we focus on a selection-mutation algorithm with infinite population evolving in the set of integers. The weighted random walk model is closely related to this algorithm : to each path of the simple random walk is assigned a weight given by a multiplicative functional. We prove a functional large deviations principle for the weighted random walk and deduce the genetic algorithm's long time behaviour.

The last part is devoted to a new modelisation of data structures, which is a generalisation of the markovian model studied by Flajolet, Louchard ... We introduce a temporal inhomogeneity in the model thanks to the dynamic random walks studied by Guillotin and we define the dynamic model for data structures. We prove a functional large deviations principle for the size process and a functional law of large numbers.

Key words : Large Deviations ; Laplace Method ; DNA Denaturation ; Genetic Algorithm ; Data Structures ; Dynamic Random Waks.

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à mes directeurs de thèse, Christian Mazza et Nadine Guillotin-Plantard. Je remercie Christian Mazza pour m'avoir fait découvrir des problèmes situés à l'interface des probabilités et de la biologie pour lesquels la théorie n'est pas encore écrite. Son intuition, son enthousiasme et sa bonne humeur m'ont beaucoup apporté. Je suis reconnaissant à Nadine Guillotin-Plantard qui m'a ouvert d'autres champs de recherche en m'initiant aux marches aléatoires dynamiques. J'ai grandement profité de ses compétences, de sa disponibilité, de ses conseils et de son soutien, et je l'en remercie chaleureusement.

J'adresse également mes plus sincères remerciements à Françis Comets et Craig John Benham pour l'honneur qu'ils me font en s'intéressant à mes travaux, en acceptant d'être les rapporteurs de cette thèse.

Stéphane Attal a toute ma gratitude pour avoir accepté de faire partie de ce jury. Sa gentillesse et son dynamisme contribuent grandement à la bonne ambiance au sein de l'équipe *analyse stochastique* du laboratoire.

Je souhaite également remercier ici Alice Guyonnet, André Goldman et Didier Piau, qui sont des modèles de compétence et qui ont nourri par leur cours et exposés remarquables mon goût pour les probabilités. Je pense également à Arnaud Le Ny, Laurent Gueguen, Renée Schott, Jean Bérard, Pierre Pudlo, Véronique Ladret et Pierre Bousquet qui ont été des interlocuteurs de grande valeur. Nos discussions fructueuses m'ont permis de prendre du recul sur mon travail, de résoudre certaines difficultés, en un mot, d'avancer.

Enfin, tout mes collègues, de Gerland et de la Doua ont participé à créer une ambiance de travail agréable et détendue. Je salue particulièrement Anne, Gabriela, Frédérique, Jean-Baptiste, Mariam, Pierre, Stéphane, Clément, Ricardo, Jérémie, ainsi que tous les membres du bureau 111.

Table des matières

Aperçu sur le contenu et l'organisation de ce document

Outre une assez longue introduction visant à présenter nos résultats en français et à les replacer dans leur contexte, cette thèse comporte quatre chapitres. Ces différents chapitres sont largement indépendants, afin que le lecteur puisse prendre la lecture de ce document où bon lui semble.

Le premier chapitre constitue une introduction rapide à la théorie des grandes déviations. Cette théorie a connu de nombreux développements et applications, il s'agit ici d'en présenter les concepts fondamentaux. L'approche est inspirée de l'ouvrage "Large deviations techniques and applications" de Dembo et Zeitouni [19]. Les preuves sont généralement omises. Une exception est faite pour la méthode de Laplace (connue aussi sous le nom de lemme intégral de Varadhan), qui constitue un outil central de nos travaux. Parmi les nombreux résultats classiques de grandes déviations, nous présentons ceux qui nous ont servi : les principes de grandes déviations dans $\mathbb{R}^d$ avec le théorème de Gärtner-Ellis, les principes fonctionnels de grandes déviations avec le théorème de Mogulskii.

Le deuxième chapitre reprend l'article [21] "A stochastic model for DNA denaturation" publié dans *Journal of Statistical Physics*. Ce travail est dans la lignée des travaux de Benham [2, 3, 4] et de Mazza [45] qui ont modélisé la dénaturation de l'ADN en termes de physique statistique. Nous reprenons le modèle proposé par Mazza et parvenons à expliciter son comportement dans une assez grande généralité, c'est à dire en prenant en compte l'hétérogénéité des bases constituant le polymère d'ADN. La capacité de rendre compte cette hérérogénéité dans des résultats théoriques est un trait particulièrement satisfaisant de cette étude. Nous obtenons des résultats de type loi des grands nombres grâce à la théorie des grandes déviations via la méthode de Laplace. Ces résultats permettent de décrire la dénaturation de l'ADN en fonction de certains paramètres, notamment la température et la superhélicité.

Dans le troisième chapitre, nous reprenons l'article [22] "A weighted random walk model - Application to a genetic algorithm" soumis à *Journal of Applied*

Probability. Nous présentons la marche pondérée, obtenue à partir de la marche simple sur $\mathbb{Z}$ en pondérant les trajectoires grâce à une fonctionnelle multiplicative. Nous obtenons via des techniques de grandes déviations une loi des grands nombres fonctionnelle. Des difficultés techniques dans l'application de la méthode de Laplace apparaissent à cause de l'irrégularité des fonctionnelles en jeu. La résolution du problème variationnel permet d'expliciter la limite déterministe de la marche pondérée correctement renormalisée. Nous proposons enfin une application de ces résultats à l'étude de la rapidité de la fuite vers l'infini d'un algorithme génétique, généralisant ainsi des résultats de Mazza et Piau [46].

Le quatrième chapitre est issu de l'article [23] "Data Structures with Dynamical Random Transitions" écrit en collaboration avec N.Guillotin-Plantard, B.Pinçon et R.Schott, accepté pour publication dans *Random Structures and Algorithms*. Il s'agit de présenter une nouvelle modélisation des structures de données (piles, listes linéaires, queues de priorité, dictionnaires ...). L'idée principale est d'introduire une hétérogénéité temporelle dans des modèles plus anciens étudiés entre autre par Louchard [41]. La marche simple est remplacée par la marche dynamique introduite par N.Guillotin-Plantard [30, 29]. Nous obtenons des comportements de type loi des grands nombres pour l'évolution de la taille des structures de données dynamiques. Le coût de stockage est également étudié.

Enfin, une dernière partie présente quelques questions restées ouvertes que nos travaux ont soulevées. Certaines ont donné lieu à une recherche relativement conséquente, mais nous ne sommes pas parvenus à trouver de réponses satisfaisantes. D'autres n'ont pas encore été étudiées, et constituent des bases pour nos recherches à venir.

Introduction

Introduction générale

A première vue, cette thèse pourrait sembler quelque peu hétérogène. Plusieurs thèmes très différents y sont traités, à priori sans rapport les uns avec les autres :

- la biologie moléculaire, avec l'étude de la dénaturation de l'ADN (c'est-à-dire la façon dont les brins de l'ADN se séparent, permettant d'assurer les fonctions de duplication et d'expression du génome),
- la théorie des algorithmes génétiques, théorie mathématique visant à résoudre des problèmes d'optimisation discrète,
- l'informatique théorique, avec l'étude de différents types de structures de données.

La cohérence de ce travail réside dans la formulation mathématique et la résolution des problèmes soulevés. En effet, pour traiter ces différents sujets, le mathématicien commence par une phase de modélisation : comment peut-on représenter en termes mathématiques le phénomène que l'on étudie ? quelles sont les grandeurs d'intérêt ?
Il s'avère que dans ces différents domaines, la théorie des probabilités est un outil de modélisation pertinent et efficace. Pour modéliser le phénomène, on décrit un espace d'états, où chaque état correspond à une certaine configuration possible du système étudié. Puis on munit cet espace d'une mesure de probabilité, qui permet une description des configurations les plus ou moins fréquentes. Une fois ce modèle fourni, le mathématicien peut s'attacher à étudier les mesures de probabilité introduites et les variables aléatoires représentant les grandeurs d'intérêt.

Un calcul direct est cependant souvent délicat, voire hors d'atteinte. Nous nous penchons alors sur le comportement asymptotique du système quand sa taille devient très grande. Par exemple, dans l'étude de la dénaturation de l'ADN, les polymères d'ADN ont une longueur de l'ordre de 10^3 à 10^6 paires de bases. Pour l'étude des bases de données, nous nous intéressons à leur évolution sur le long terme, le temps d'observation est grand. Lorsque la taille du

système grandit, des phénomènes de moyennisation apparaissent, et on observe un comportement quasi-déterministe de certaines quantités correctement renormalisées. Ainsi, dans un jeu de pile ou face d'une durée relativement longue, la proportion de pile obtenue est très proche de $0,5$. Ce comportement quasi déterministe des grands systèmes constitue ce qu'on appelle la *loi des grands nombres*. Dans cette thèse, nous obtenons pour les différents modèles étudiés des résultats de type loi des grands nombres, c'est-à-dire que nous montrons que les principales grandeurs étudiées ont un comportement quasi-déterministe lorsque la taille du système grandit. Ces comportements limites sont également étudiés et caractérisés.
Pour obtenir ces résultats, nous utilisons de manière essentielle la théorie des grandes déviations. Cette théorie permet de quantifier la façon dont le comportement du système diffère de son comportement typique (celui donné par une loi des grands nombres par exemple). Elle donne un équivalent dans une échelle logarithmique pour la probabilité qu'à le système de dévier de son comportement typique. Nous commençons par étudier les propriétés de type grandes déviations des différents systèmes et en déduisons des résultats de type loi des grands nombres. Nous utilisons la méthode de Laplace pour obtenir différents principes de grandes déviations. Cette méthode trouve des applications dans de nombreux domaines, où la théorie des grandes déviations intervient de manière naturelle (par exemple, le lien entre la thermodynamique et la théorie des grandes déviations est très bien expliqué dans l'article de Lewis et Pfister "Thermodynamic probability theory : Some aspects of large deviations" [39], ou pour une approche via la mécanique statistique, on consultera le livre de Ellis "Entropy, Large Deviations, and Statistical Mechanics" [24]).

Cette thèse contient donc un travail de modélisation pour formaliser des questions provenant de différents domaines dans le langage des probabilités. Puis nous nous servons de l'outil des grandes déviations pour étudier les modèles probabilistes introduits. Grâce à la méthode de Laplace, nous obtenons des principes de grandes déviations et en déduisons des résultats de type lois des grands nombres. Les comportements limites sont identifiés en résolvant les problèmes variationnels de minimisation des fonctions de taux.

Après avoir présenté la problématique générale et les méthodes utilisées, nous introduisons maintenant les différents thèmes abordés, les modèles utilisés ainsi que les principaux résultats obtenus.

Introduction au chapitre 2 "A stochastic model for DNA denaturation"

Motivations

En 1953, Watson et Crick ont mis en évidence la structure en double hélice de

l'ADN. Cette découverte a marqué les esprits car on a alors pu se représenter les mécanismes de la vie au niveau moléculaire, en particulier les mécanismes de réplication et de transcription. La réplication de l'ADN consiste à dupliquer le matériel génétique et joue un rôle essentiel lors la division cellulaire. La transcription permet l'expression des gènes : la séquence de nucléotides est convertie en séquence d'acides aminés selon le code génétique afin de synthétiser les protéines. Dans ces processus, le fait que les deux brins de l'ADN puissent se séparer est capital. On appelle dénaturation de l'ADN la séparation des deux brins, qui peut être partielle (lors de la transcription) ou bien totale (lors de la réplication). Lors de la dénaturation de l'ADN, les liaisons hydrogènes entre les nucléotides complémentaires sont rompues. Nous nous proposons ici d'étudier un modèle de dénaturation de l'ADN.

Modélisation

Notre approche se place dans la lignée des travaux de Benham ([2, 3, 4]). Le modèle initial est un modèle de spins en dimension 1 et en champ hétérogène. Un graphe circulaire de taille N représente un polymère d'ADN constitué de N paires de nucléotides. Il y a quatre types de nucléotide : adénine, thymine, cytosine, guanine notés respectivement A,T,C,G. Ces nucléotides ont la particularité de pouvoir se lier entre eux par des liaisons hydrogène, en respectant certaines règles de complémentarité : d'une part A et T peuvent former des liaisons, et d'autre part G et C. Le polymère d'ADN est constitué d'une double chaîne de nucléotides. Les deux chaînes sont complémentaires, et au niveau de chaque paire, les liaisons hydrogènes peuvent être ouvertes ou fermées. Chaque sommet du graphe représente une paire de nucléotides, soit A-T, soit G-C. La force de la liaison dépend des nucléotides en présence, la liaison GC étant plus forte que la liaison AT. On note b_i^N l'énergie de liaison au site i qui peut prendre la valeur b_{AT} ou b_{GC}. Dans notre modèle, la séquence $B_N = (b_i^N)_{1\leq i\leq N}$ joue le rôle d'un champ extérieur à valeur dans $\{b_{AT}, b_{GC}\}^N$. Au niveau du sommet i, la liaison peut être fermée (ce que l'on représente par un spin $\sigma_i = +1$) ou ouverte (correspondant à un spin $\sigma_i = -1$). L'état des différentes liaisons est représenté par une configuration $\sigma = (\sigma_i)_{1\leq i\leq N} \in \Omega_N$, où $\Omega_N = \{-1, +1\}^N$ est l'espace des configurations. Par commodité, on introduit les variables

$$n_i = \frac{1+\sigma_i}{2} = \begin{cases} +1 & \text{si} \quad \sigma_i = +1 \\ 0 & \text{si} \quad \sigma_i = -1 \end{cases}, \quad i = 1, \cdots, N.$$

La proportion de liaisons ouvertes est donnée par

$$M_N(\sigma) = \frac{1}{N}\sum_{i=1}^{N} n_i \in [0, 1].$$

Cette quantité est une mesure de la dénaturation : pour $M_N = 0$, toutes les liaisons sont fermées, la dénaturation est nulle tandis que pour $M_N = 1$, toutes les liaisons sont ouvertes, la dénaturation est totale. Dans la terminologie du

modèle d'Ising, $M_N(\sigma)$ est l'aimantation.
L'énergie totale de liaison est donnée, au facteur $1/N$ près, par la somme

$$M_{NB_N}(\sigma) = \frac{1}{N}\sum_{i=1}^{N} b_i^N n_i.$$

Enfin, le nombre de "poches de dénaturation" (qui dans la terminologie du modèle d'Ising correspond au périmètre) est défini par la formule

$$R_N(\sigma) = \frac{1}{2}\sum_{i=1}^{N} \mathbb{1}_{\sigma_i \neq \sigma_{i+1}},$$

où on utilise la condition aux bords périodique $\sigma_{N+1} = \sigma_1$.

Benham a introduit l'hamiltonien

$$H_N(\sigma) = aR_N(\sigma) + NF(M_N(\sigma), M_{NB_N}(\sigma))$$

où F est la fonction rationnelle

$$F(m, \tilde{m}) = \frac{2\pi^2 C K_0}{4\pi^2 C + K_0 m}(\kappa + \frac{m}{A})^2 + \tilde{m}.$$

Cet hamiltonien prend en compte :
– l'énergie nécessaire pour amorcer la dénaturation (à travers la constante a),
– l'énergie des liaisons AT et GC (à travers les constantes b_{AT} et b_{GC}),
– des énergies liées à l'élasticité du polymère d'ADN et à sa géométrie : l'énergie de torsion (à travers la constante C) et l'énergie associée au linking number résiduel (à travers la constante K_0). Ces termes énergétiques font intervenir la superhélicité κ du polymère.

Les valeurs numériques de ces constantes ont été évaluées expérimentalement par Sun,Mezei, Fye and Benham [50].

La mesure de Gibbs sur l'espace des configurations Ω_N construite à partir de cet hamiltonien est

$$\pi_N(\sigma) = \frac{1}{Z_N} e^{-H_N(\sigma)} \rho_N(\sigma)$$

où ρ_N désigne la mesure uniforme sur Ω_N et Z_N la fonction de partition

$$Z_N = \int_{\Omega_N} e^{-H_N(\sigma)} \rho_N(d\sigma).$$

Dans ce modèle, étudier la dénaturation revient à étudier le comportement de la variable aléatoire $M_N : \Omega_N \to [0, 1]$ sous la mesure de Gibbs π_N.

Travaux antérieurs

Outre des discussions importantes portant sur la modélisation, les travaux de Benham portent sur l'évaluation des termes dominants dans la fonction de

partition ainsi que sur des méthodes algorithmiques visant à localiser les régions où la dénaturation est hautement probable. Dans [45], Mazza calcule la limite thermodynamique du modèle de Benham dans le cas d'un homopolymère (i.e. en champ extérieur homogène) et obtient un principe de grandes déviations pour M_N en utilisant le théorème de Gärtner-Ellis. Il propose également une modification du modèle de Benham avec un conditionnement aux configurations ayant un nombre de "poches de dénaturation" restreint. Ceci est justifié par les biologistes qui observent que dans le processus de dénaturation, la séparation des deux brins a lieu en un nombre limité de sites puis ensuite se propage, un petit peu comme l'ouverture d'une fermeture éclair. Formellement, il introduit la mesure conditionnée

$$\pi_{Nr_N} = \pi_N(\ .\ \mid R_N \leq r_N),$$

où r_n est une suite d'entiers vérifiant $r_N = o(N)$. Pour ce nouveau modèle, il traite le cas des homopolymères ainsi qu'un exemple particulier d'hétéropolymère, et obtient des principes de grandes déviations pour la dénaturation M_N ainsi que des résultats de type loi des grands nombres. Il apparaît que dans le modèle original de Benham, les états de non dénaturation ou de dénaturation totale ne sont pas stables (ce qui se traduit par une dénaturation limite comprise strictement entre 0 et 1), cela contredit les observations biologiques. Dans le modèle modifié, ces états sont stables, et l'on passe d'un état à l'autre en faisant varier différents paramètres comme la température ou la superhélicité, avec des phénomènes de seuil. D'un point de vue qualitatif, cela correspond bien aux observations biologiques.

Contributions personnelles

Nos travaux ont permis d'étendre les résultats de Mazza à des hétéropolymères plus généraux. Nous avons une approche asymptotique du problème, donc nous travaillons avec une suite de champ $(B_N)_{N\geq 1}$ représentant une suite de polymères d'ADN de taille $N \geq 1$ et construisons la suite de mesures de Gibbs $(\pi_{Nr_N})_{N\geq 1}$. Afin de dégager une asymptotique, il faut évidemment que la suite de champs B_N soit cohérente, qu'elle converge en un certain sens. Dégager les hypothèses adéquates sur le champ extérieur B_N constitue une part significative du travail. Une fois cela mis en place, nous prouvons un principe de grande déviations pour le triplet $(M_N, M_{NB_N}, R_N/N)$ sous la mesure π_{Nr_N}.

Pour construire la suite de mesures $(\pi_{Nr_N})_{N\geq 1}$, on se donne une suite d'entiers r_N et une suite de champs $B_N \in \{b_{AT}, b_{GC}\}^N$. Détaillons les hypothèses utilisées. On se place dans l'hypothèse de petit périmètre $R_N \leq r_N$, où r_N vérifie

- $(\mathbf{H_0})$: $\lim\limits_{N\to+\infty} \frac{r_N}{N} = 0.$

Les hypothèses sur le champ extérieur portent sur une moyenne spatiale locale

du champ notée $\overline{B}_N = (\bar{b}_i^N)_{1\leq i\leq N} \in [b_{AT}, b_{GC}]^N$ et définie par

$$\bar{b}_i^N = \frac{1}{T_N} \sum_{k=i}^{i+T_N-1} b_k^N \ , \ 1 \leq i \leq N$$

où T_N est l'échelle utilisée pour faire la moyenne et où on applique la condition aux bords périodique $b_{N+i}^N = b_i^N$. On suppose que les champs moyennés sont "presque constants par morceaux, avec au plus r_N morceaux". Formellement, introduisons L_{Nr_N} comme l'ensemble des champs $G = (g_i)_{1\leq i\leq N} \in \mathbb{R}^N$ défini de la manière suivante : $G \in L_{Nr_N}$ si et seulement si il existe des entiers $0 = l_0 < l_1 < \cdots < l_{r_N} = N$ tels que g_i est constant sur chacun des r_N intervalles de la forme $l_k < i \leq l_{k+1}$. On suppose que les champs moyennés $\overline{B}_N$ sont "presque" dans L_{Nr_N} au sens suivant :

• $(\mathbf{H_1})$: Pour tout $N \geq 1$, il existe un champ $\widetilde{B}_N = (\tilde{b}_i^N)_{1\leq i\leq N} \in L_{Nr_N}$ tel que la distance $\delta_N := \frac{1}{N}\sum_{i=1}^N |\bar{b}_i^N - \tilde{b}_i^N|$ tende vers zero lorsque $N \to +\infty$.

On impose d'autre part une convergence des champs de la façon suivante :

• $(\mathbf{H_2})$: La suite de mesures de probabilité $\mu_{\overline{B}_N} := \frac{1}{N}\sum_{i=1}^N \delta_{\bar{b}_i^N}$ converge en loi vers une mesure de probabilité μ lorsque $N \to +\infty$.

Enfin, on impose une condition sur les différentes échelles en jeu

• $(\mathbf{H_3})$: $\lim_{N\to+\infty} \frac{r_N T_N}{N} = 0$.

Cela signifie que l'échelle de moyennisation T_N est petite devant la longueur moyenne d'un domaine de dénaturation N/r_N. Dans le chapitre 2, le lecteur trouvera une discussion de ces différentes hypothèses, différents exemples de champs les vérifiant (dont des modèles de champs aléatoires) ainsi qu'une interprétation de ces hypothèses du point de vue du biologiste.

Le principal résultat est un principe de grandes déviations (PGD) pour le triplet $(M_N, M_{NB_N}, R_N/N)$ sous la mesure π_{Nr_N} :

Théorème 0.1. *On suppose les hypothèses $(H_0) - (H_3)$ vérifiées. Alors la loi du triplet $(M_N, M_{NB_N}, R_N/N)$ sous la mesure π_{Nr_N} vérifie un PGD de vitesse N et bonne fonction de taux J_Δ définie par*

$$J_\Delta(m, \tilde{m}, r) = \begin{cases} F(m, \tilde{m}) - \inf_\Delta F & \text{si } (m, \tilde{m}) \in \Delta \text{ et } r = 0 \\ +\infty & \text{sinon} \end{cases}$$

où Δ est le domaine

$$\Delta = \left\{ (m, \tilde{m}) \in \mathbb{R}^2 \mid 0 \leq m \leq 1 \ , \ \int_0^m F_\mu^{-1}(x)\, dx \leq \tilde{m} \leq \int_{1-m}^1 F_\mu^{-1}(x)\, dx \right\}$$

et où F_μ^{-1} désigne la pseudo-inverse de la fonction de répartition de la mesure de probabilité μ.

La preuve de ce théorème repose sur un PGD pour le triplet sous la mesure uniforme sur l'ensemble $\{\sigma \in \Omega_N \mid R_N(\sigma) \leq r_N\}$. La fonction de taux est dégénérée, égale à zero sur $\Delta \times \{0\}$ et à $+\infty$ en dehors. Cela traduit le fait que le support de la loi du triplet est asymptotiquement $\Delta \times \{0\}$, ce que nous prouvons par une étude combinatoire du système (plus précisément, nous réduisons la preuve à un lemme combinatoire établissant l'inexistence ou l'existence de certaines configurations). Une application directe de la méthode de Laplace permet ensuite de prouver le théorème précédent.

Nous en déduisons ensuite une loi des grands nombres :

Proposition 0.1. *Soit $\kappa \neq \frac{4\pi^2 C}{K_0 A}$. La loi de $(M_N, \widetilde{M}_{NB_N}, R_N/N)$ sous π_{Nr_N} converge lorsque N tend vers $+\infty$ vers la mesure de Dirac au point $(M_\infty, \widetilde{M}_\infty, 0)$. Ce point est l'unique minimiseur de la fonction F sur Δ et satisfait*

$$\widetilde{M}_\infty = \int_0^{M_\infty} F_\mu^{-1}(x)\, dx.$$

Ce résultat donne ensuite lieu à une interprétation biologique et à une illustration numérique.

Introduction au chapitre 3
" A weighted random walk model - Application to a genetic algorithm"

Motivations

Les algorithmes génétiques en population finie ont été introduits par Holland en 1975 [33] et ont depuis trouvé de nombreuses applications. On les retrouve dans de nombreux domaines, par exemple la biologie, l'informatique, la combinatoire ... Typiquement, ces algorithmes sont utilisés dans des problèmes d'optimisation difficiles, où il faut trouver les valeurs maximales d'une fonction sur un ensemble plus ou moins complexe. Une population d'individus, considérés comme des solutions éventutelles au problème, évolue sous l'influence de deux phénomènes : les mutations qui vont permettre à la population d'explorer l'espace ambiant, et la sélection qui va favoriser les meilleurs individus pour le problème considéré. Certains modèles prennent également en compte un phénomène de reproduction ("crossing-over") que nous ne développerons pas ici. La dynamique d'évolution simule, à l'instar des systèmes naturels, la survie des

individus les mieux adaptés. Les mécanismes de la dynamique sont inspirés des mécanismes génétiques de la vie, d'où le nom "algorithmes génétiques".

Modélisation
D'un point de vue mathématique, les algorithmes génétiques sont des chaînes de Markov sur un espace produit E^p, où E est l'espace des configurations et $p \geq 1$ est la taille de la population. La dynamique est une combinaison de deux opérateurs de base : mutation et sélection. La mutation opère par un noyau de transition ergodique $Q(.,.)$ sur E. Soit $x = (x_i)_{1 \leq i \leq p} \in E^p$ une population. L'effet de la mutation sur x est modélisé par le choix aléatoire d'une nouvelle population avec probabilité

$$Q(x_1, .) \otimes Q(x_2, .) \otimes \cdots \otimes Q(x_p, .).$$

C'est à dire que lors de la mutation chaque individu effectue une transition selon le noyau $Q(.,.)$ indépendamment des autres individus. La sélection utilise une fonction $f \geq 0$ sur E, appelée fonction d'adaptation (ou fitness), qui est usuellement la fonction que l'on cherche à rendre maximale. Etant donné une population x, on définit la mesure de probabilité π_x sur $\{x_1, \cdots, x_p\}$ par

$$\pi_x = \frac{1}{< f(x) >} \sum_{i=1}^{p} f(x_i) \delta_{x_i}, \quad < f(x) >= \sum_{i=1}^{p} f(x_i).$$

La sélection consiste à choisir la nouvelle population dans l'ensemble $\{x_1, \cdots, x_p\}^p$ selon la mesure de probabilité $\pi_x^{\otimes p}$. Notons que cette probabilité favorise les individus correspondant aux grandes valeurs de la fonction d'adaptation f. Une population initiale étant donnée X_0, l'algorithme la fait évoluer au cours du temps en lui faisant subir successivement et alternativement les opérations de mutation et de sélection :

$$X_n \overset{mutation}{\longrightarrow} Y_n \overset{sélection}{\longrightarrow} X_{n+1}.$$

L'idée heuristique de l'algorithme génétique est de permettre aux individus d'explorer l'espace d'états E grâce aux mutations, et de favoriser les mutations favorables (i.e. celles qui donnent des individus avec un grand fitness) grâce aux sélections. On espère ainsi que la population va évoluer vers les individus ayant un fitness maximum, permettant ainsi de résoudre le problème d'optimisation de départ.

Travaux antérieurs
Dès leur découverte, les algorithmes génétiques ont connus un succès relativement grand dû principalement à la facilité de leur mise en œuvre et à leur capacité d'être utilisés dans de nombreux cas avec de bons résultats. De nombreuses études expérimentales et heuristiques leur ont été consacrées, donnant lieu à un savoir-faire empirique concernant la manière d'ajuster les différents

paramètres au modèle considéré. Mais malgré leur efficacité pratique, les résultats théoriques décrivant le comportement des algorithmes génétiques sont assez rares. En effet, il s'agit d'étudier des chaînes de Markov en grande dimension (la dimension grandissant très rapidement avec la taille de la population), et du fait de la sélection, la dynamique est généralement irréversible. Les premiers résultats de convergence asymptotique ont été obtenus par R. Cerf [11] [12] dans le cadre de la théorie de Freidlin-Wentzell des chaînes de Markov à transitions rares. Il s'agit d'une théorie perturbative dans laquelle l'intensité de la sélection tend vers l'infini et le taux de mutation tend vers zéro. Un autre cadre asymptotique dans lequel des résultats ont été obtenus est celui où la taille de la population est supposée infinie. Plus précisément, on peut montrer sous des hypothèses très générales que lorsque la taille de la population p tend vers l'infini, la mesure empirique converge en loi vers une mesure de probabilité limite μ_n, pour tout instant n fixé. L'intérêt de se placer dans ce contexte est que les fluctuations stochastiques liées à la taille finie de la population disparaissent. La suite de mesures μ_n forme un système dynamique non linéaire à valeurs mesures :

$$\mu_{n+1} = T(\mu_n).$$

L'opérateur $T = W \circ M$ est la composée de l'opérateur de mutation M et de l'opérateur de sélection W. L'opérateur de mutation est défini par M : $\mu \mapsto \mu Q$, où Q est le noyau markovien de mutation. La sélection agit de la manière suivante : pour une mesure de probabilité μ telle que $\int f d\mu \in (0, +\infty)$, l'opérateur de sélection W remplace μ par $W(\mu) = \hat{\mu}$ défini par

$$\frac{d\hat{\mu}}{d\mu}(x) = \frac{f(x)}{\int f d\mu}$$

Les propriétés asymptotiques de ce système ont notamment été étudiées par M.Vose [56], Y. Rabinovich and A. Wigderson [49] et également C. Mazza and D. Piau [46]. D'autre part, P. Del Moral and A. Guionnet [16, 17, 18] ont étudié du point de vue des grandes déviations la convergence des mesures empiriques en population finie vers les mesures limite, en lien également avec des questions de filtrage non-linéaire. Dans nos travaux, nous étudions un exemple particulier d'algorithme de sélection-mutation : l'espace des états est $E = \mathbb{Z}$, la mutation correspond à un pas de la marche simple symétrique, le noyau de transition étant donné par

$$Q(x, y) = \begin{cases} \frac{1}{2} & \text{si } |x - y| = 1 \\ 0 & \text{sinon.} \end{cases}$$

Ce modèle d'évolution a été introduit en biologie pour modéliser l'évolution d'une population de virus [53, 34]. J. Bérard and A. Bienvenüe [5, 6] ont donné une asymptotique fine du comportement de cet algorithme, avec notamment un théorème d'invariance. Nos travaux se placent dans la continuité d'un travail de C.Mazza et D.Piau [46] où le modèle est étudié en population infinie pour la fonction fitness linéaire $f(x) = x$. Les auteurs y décrivent précisément le comportement asymptotique en temps long du système dynamique

avec notamment une loi des grands nombres, un théorème central limite et un principe de grandes déviations. Leur preuve est basée sur une estimation précise des transformées de Laplace.

Contributions personnelles
Nos travaux étendent les résultats obtenus par Mazza et Piau [46] et permettent de prendre en compte d'autres fonctions fitness (les fonctions puissances) et également d'autres types de mutation.

Théorème 0.2. *Soit $\beta > 0$. Pour l'algorithme génétique en population infinie sur les entiers avec mutation basée sur la marche simple symétrique et fonction fitness $f(x) = x^\beta$, la population au temps n est située près du point $v_\beta n$. Plus précisément, si X_n a pour loi μ_n, alors $\frac{X_n}{n}$ converge presque sûrement vers v_β. La vitesse est*

$$v_\beta = \left(\int_0^1 \frac{dx}{\sqrt{1 - x^{2\beta}}} \right)^{-1}.$$

Les techniques employées sont relativement différentes : une étude directe des transformées de Laplace étant impossible, nous avons reformulé le problème en terme de marche aléatoire et introduit le modèle de marche aléatoire pondérée. Nous nous servons alors de la théorie des grandes déviations et de la méthode de Laplace dans un cadre fonctionnel.

Soit Ω_n l'ensemble des trajectoires de longueur n

$$\Omega_n = \left\{ S = (S_0, \cdots, S_n) \in \mathbb{Z}^{n+1} \mid \forall k = 1, \cdots, n,\ |S_k - S_{k-1}| = 1 \right\}.$$

Soit $\mathbb{P}_{\pi,n}$ la mesure de probabilité sur Ω_n correspondant à la marche simple symétrique de loi initiale π. Définissons le poids d'une trajectoire $S \in \Omega_n$ par

$$\Pi_n^f(S) = \prod_{k=1}^{n} f(S_k).$$

Nous introduisons le modèle de marche pondérée avec la mesure de probabilité $\mathbb{P}_{\pi,n}^f$ sur Ω_n définie par

$$\mathbb{P}_{\pi,n}^f(S) = \frac{1}{Z_{\pi,n}^f} \Pi_n^f(S) \mathbb{P}_{\pi,n}(S)$$

où $Z_{\pi,n}^f$ est la fonction de partition

$$Z_{n,\pi}^f = \int_{\Omega_n} \Pi_n^f(S) d\mathbb{P}_{\pi,n}(S).$$

que nous supposons non nulle. Nous utilisons la notation $\mathbb{P}_{\pi,n}^\beta$ lorsque la fonction fitness est la fonction puissance $f(x) = x^\beta$, $\beta > 0$.

Le modèle de marche pondérée et l'algorithme génétique en population infinie sont reliés par le résultat suivant : μ_n est la loi de S_n sous la mesure $\mathbb{P}^f_{\mu_0,n}$. Nous déduirons donc les propriétés asymptotiques de l'algorithme génétique de celles de la marche pondérée. Les résultats que nous avons obtenus pour la marche pondérée sont les suivants. Notons Ω l'ensemble des fonctions càd-làg sur $[0,1]$, et renormalisons la marche pondérée grâce aux applications $\Psi_n : \Omega_n \to \Omega, S \mapsto \Psi_n(S)$ définies par

$$\Psi_n(S) : t \mapsto \begin{cases} \frac{1}{n} S_{[nt]+1} & \text{si } 0 \leq t < 1 \\ \frac{1}{n} S_n & \text{si } t = 1 \end{cases} .$$

Notons $\widetilde{\mathbb{P}}^f_{\pi,n} = \mathbb{P}^f_{\pi,n} \circ \Psi_n^{-1}$.

Théorème 0.3. *Soit $\beta > 0$ fixé.*

1. *La suite de mesures de probabilité $\widetilde{\mathbb{P}}^\beta_{\pi,n}$ vérifie la borne supérieure de grandes déviations suivante : pour tout ensemble fermé $A \subset \Omega$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \widetilde{\mathbb{P}}^\beta_{\pi,n}(A) \leq - \inf_{\phi \in A} \left\{ I - \beta F - \inf_{\phi \in \Omega} (I - \beta F) \right\}$$

où les fonctionnelles $F : \Omega \longrightarrow [-\infty, +\infty[$ et $I : \Omega \longrightarrow [0, +\infty[$ sont définies par :

$$F(\phi) = \int_0^1 \log(\phi(t))\ dt,$$

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^\star(\dot{\phi}(t))\ dt & si\ \phi \in \mathcal{AC}\ et\ \phi(0) = 0 \\ +\infty & sinon \end{cases},$$

$\mathcal{AC}$ étant l'ensemble des fonctions absolument continues sur $[0,1]$ et $\Lambda^\star : \mathbb{R} \to [0, +\infty]$ la fonction de taux associée à la marche simple symétrique

$$\Lambda^\star(x) = \begin{cases} \frac{1-x}{2} \log(1-x) + \frac{1+x}{2} \log(1+x) & si\ x \in [-1,1] \\ +\infty & sinon \end{cases} .$$

2. *La fonctionnelle $I - \beta F$ a un unique minimiseur sur Ω noté ψ_β. La suite de mesures de probabilité $\widetilde{\mathbb{P}}^\beta_{\pi,n}$ converge vers δ_{ψ_β} avec une vitesse exponentielle. C'est-à-dire que pour tout $\varepsilon > 0$, il existe $\delta = \delta(\varepsilon) > 0$ tel que pour tout n suffisamment grand,*

$$\widetilde{\mathbb{P}}^\beta_{\pi,n} \left(\|\phi - \psi_\beta\|_\infty > \varepsilon \right) \leq e^{-n\delta}.$$

3. *Soit G_β la fonction définie par*

$$G_\beta(y) = \int_0^y \frac{dx}{\sqrt{1 - x^{2\beta}}}.$$

La fonction ψ_β est égale à

$$\psi_\beta(t) = \frac{1}{G_\beta(1)} G_\beta^{-1} \left(G_\beta(1) t \right).$$

Pour prouver la borne supérieure de grandes déviations, nous utilisons le théorème de Mogulskii qui donnent les propriétes de grandes déviations de la marche simple. La méthode de Laplace permet de passer aux propriétés de grandes déviations pour la marche pondérée. Cependant, l'irrégularité de la fonctionnelle F qui n'est ni bornée ni continue empêche une application directe de la méthode et nous devons regarder de plus près ce qu'il se passe (notamment au voisinage du point 0) pour obtenir cette borne supérieure.

Introduction au chapitre 4 "Data structures with dynamical random transition"

Motivations

Les structures de données sont une notion de bases en informatique théorique. Il s'agit de gérer un ensemble fini d'éléments dont le nombre n'est pas fixé à priori. Les éléments de cet ensemble peuvent être de différentes sortes : nombre entiers ou réels, chaînes de caractères, ou des objets informatiques plus complexes comme les identificateurs de processus ou les expressions de formules en cours de calcul ... On ne s'intéresse pas aux éléments de l'ensemble en question mais aux opérations que l'on effectue sur cet ensemble. Les structures de données sont utilisées de façon très intensive en programmation ou en gestion des bases de données. Il existe différents types de structures de données (tableaux, arbres ...), nous aborderons ici seulement les strucures élémentaires de type liste. Nous nous intéressons principalement à l'évolution de la taille de la structure.

Modélisation

Un type de donnée est la donnée des opérations basiques autorisées (insertion, suppression, requête ...) et des restrictions d'accès éventuels. Les exemples de bases sont les piles, les listes linéaires, les queues de priorité et les dictionnaires. Par exemple, dans une pile, l'insertion et la suppression se font en haut de la pile. Un schéma de longueur n est un mot $\omega = O_1 O_2 \cdots O_n$ composé des lettres I, D, Q^+ et Q^- tel que :

$$\forall 1 \leq k \leq n \ , \ |O_1 O_2 \cdots O_k|_D \leq |O_1 O_2 \cdots O_k|_I,$$

où $|O_1 O_2 \cdots O_k|_D$ (respectivement $|O_1 O_2 \cdots O_k|_I$) est le nombre de supressions (respectivement d'insertions) parmi les k premières opérations. Le mot représente la suite des opérations insertion I, suppression D, requête positive Q^+ et négative Q^- qui ont lieu. La taille de la structure après la k-ième opération est

$$\alpha_k(\omega) = |O_1 O_2 \cdots O_k|_I - |O_1 O_2 \cdots O_k|_D,$$

Dans un schéma, les cellules sur lesquelles les opérations portent ne sont pas représentées. On en tient compte dans la notion d'histoire interne de la structure qui est définit formellement de la manière suivante : une histoire interne

est une suite de la forme

$$h = O_1(r_1)O_2(r_2)\cdots O_n(r_n)$$

où $\omega = O_1O_2\cdots O_n$ est un schéma et les r_k des entiers vérifiant

$$1 \leq r_k \leq poss(O_k, \alpha_{k-1}(\omega)).$$

L'entier r_k représente la cellule sur laquelle l'opération O_k porte. La fonction de possibilité *poss* dépend du type de donnée considéré et donne le nombre de cellules sur lesquelles l'opération O peut agir lorsque la structure a une taille α. Dans le modèle markovien, la fonction de possibilité est donnée par le tableau suivant :

Type	$poss(I,\alpha)$	$poss(D,\alpha)$	$poss(Q^+,\alpha)$	$poss(Q^-,\alpha)$
Pile	1	1	0	0
Liste linéaire	$\alpha+1$	α	0	0
Queue de priorité	$\alpha+1$	1	0	0
Dictionnaire	$\alpha+1$	α	α	$\alpha+1$

De plus, le modèle markovien fait l'hypothèse simple suivante : toutes les histoires internes ont la même probabilité. Appelons $\mathbb{P}_n$ la probabilité uniforme sur l'ensemble des schémas de longueur n. Comme le nombre d'histoires internes associées au schéma ω est

$$\prod_{k=1}^{n} poss(O_k, \alpha_{k-1}),$$

on est amené a considérer la probabilité $\mathbb{Q}_n$ définie par :

$$\mathrm{d}\mathbb{Q}_n = \frac{1}{Z_n}\prod_{k=1}^{n} poss(O_k, \alpha_{k-1})\mathrm{d}\mathbb{P}_n,$$

où Z_n est la fonction de partition

$$Z_n = \sum_{\omega}\prod_{k=1}^{n} poss(O_k, \alpha_{k-1})\mathbb{P}_n(\omega).$$

Il s'agit alors d'étudier la suite de mesures $\mathbb{Q}_n$ ainsi que les différentes grandeurs (notamment la taille maximale et le coût de stockage) sous la mesure $\mathbb{Q}_n$.

Travaux antérieurs
Le modèle markovien a été étudié par Flajolet et al. [26] par des méthodes combinatoires et l'utilisation des fonctions génératrices. Des méthodes probabilistes ont été introduites par Louchard [41] et Maier [44] afin d'étudier l'asymptotique du coût de stockage en temps et en espace lorsque la longueur de la suite d'opérations tend vers l'infini. Mais différents travaux, notamment

ceux de Knuth [35], ont montré que le modèle markovien ne correspond pas à la réalité. Un nouveau modèle basé sur une nouvelle définition des fonctions de possibilité *poss* a été proposé par Knuth et étudié par Louchard, Kenyon et Schott [42], qui ont en particulier décrit finement le comportement asymptotique des maximas. Dans le modèle de Knuth, la fonction de possibilité ne dépend pas que de la taille de la stucture et de l'opération, mais également du nombre de fois où l'opération a déjà été réalisée. Le nombre de possibilités pour la i-ème insertion I ou requête négative est égal à i quelle que soit la taille de la structure. Les fonctions de possibilité pour le modèle de Knuth sont définies dans le tableau suivant :

Type	$poss(i^{ème}\ I)$	$poss(D,\alpha)$	$poss(Q^+,\alpha)$	$poss(i^{ème}\ Q^-)$
Pile	1	1	0	0
Liste linéaire	i	α	0	0
Queue de priorité	i	1	0	0
Dictionnaire	i	α	α	i

D'autres types de données prennent en compte de nouvelles opérations (batched insertion, lazy deletion ...). Nous ne développerons pas cet aspect de la théorie, mais il semble que nos travaux puissent s'étendre également à ce type de données.

Contributions personnelles

Devant les faiblesses du modèle markovien, nous avons introduit un nouveau modèle, baptisé modèle dynamique de structures de données. Ce modèle est une généralisation du modèle markovien, mais l'hypothèse d'équiprobabilité des histoires internes est relachée en introduisant une inhomogénéité temporelle. Il est basé sur les travaux de N.Guillotin-Plantard concernant les marches aléatoires dynamiques [30, 29, 31]. L'introduction d'une inhomogénéité temporelle intéresse les informaticiens, car elle paraît naturelle dans la modélisation des phénomènes liées aux bases de données : par exemple, la fréquence des requêtes sur une base de données peut dépendre de l'heure de la journée, selon le comportement des utilisateurs. On suppose que l'inhomogénéité est due à un phénomène extérieur modélisé par un système dynamique $S = (E, \mathcal{A}, \mu, T)$, c'est-à-dire un espace probabilisé $(E, \mathcal{A}, \mu)$ muni d'une application $T : E \to E$. L'évolution du système dynamique est donnée par un point de départ $x_0 = x \in E$, puis on itère l'application T à chaque pas de temps : $x_{n+1} = T(x_n)$. La fréquence des différentes opérations dépend du système dynamique à travers une fonction mesurable $f : E \to [0,1]$. Nous présentons ici le cas des listes linéaires dynamiques et des queues de priorité dynamiques, où les opérations de base sont l'insertion I et la suppression D de cellules. Dans le chapitre 4 sont également présentés les dictionnaires dynamiques. On définit la mesure de probabilité $\tilde{\mathbb{P}}^{\star}_{n,x}$ sur l'ensemble des mots $\omega = O_1 O_2 \cdots O_n$ constitués des lettres I et D par

$$\begin{cases} O_1, O_2, \cdots, O_n \text{ sont indépendantes,} \\ \tilde{\mathbb{P}}^{\star}_{n,x}(O_k = I) = 1 - \tilde{\mathbb{P}}^{\star}_{n,x}(O_k = D) = f(T^k x). \end{cases}$$

Puis on définit $\mathbb{P}^\star_{n,x}$ par restriction sur l'ensemble des schémas avec taille initiale et finale 0 (rappelons qu'un schéma est une suite d'opérations soumise à la restriction que la taille de la structure est toujours positive). Le modèle dynamique de structure de données est alors construit par analogie avec le modèle markovien : on définit la mesure de probabilité $\mathbb{Q}^\star_{n,x}$ sur l'ensemble des schémas par

$$\mathrm{d}\mathbb{Q}^\star_{n,x} = \frac{1}{Z^\star_n} \prod_{k=1}^{n} poss(O_k, \alpha_{k-1}) \mathrm{d}\mathbb{P}^\star_{n,x},$$

où $Z^\star_n$ est la fonction de partition

$$Z^\star_n = \sum_{\omega} \prod_{k=1}^{n} poss(O_k, \alpha_{k-1}) \mathbb{P}^\star_{n,x}(\omega).$$

Lorsque $f \equiv 1/2$, on retrouve le modèle markovien : la mesure $\mathbb{P}^\star_{n,x}$ est la mesure uniforme sur l'ensemble des schémas, et la mesure $\mathbb{Q}^\star_{n,x}$ du modèle dynamique est égale à la mesure $\mathbb{Q}_n$ du modèle markovien.

Nous étudions α_k la taille du système après la k-ième opération, $0 \leq k \leq n$. Introduisons le processus de taille renormalisé $\left(\frac{1}{n}\alpha_{[nt]}\right)_{0\leq t\leq 1}$ à valeurs dans l'espace Ω des fonctions càd-làg sur $[0,1]$. Nous obtenons la borne supérieure de grandes déviations suivante.

Théorème 0.4. *Supposons que le système dynamique S est uniquement ergodique et que la fonction $\log[f(1-f)]$ est $\mu-$intégrable.*
Alors, pour tout point $x \in E$, le processus de taille renormalisé $\left(\frac{1}{n}\alpha_{[nt]}\right)_{0\leq t\leq 1}$ sous la mesure $\mathbb{Q}^\star_{n,x}$ vérifie la borne supérieure de grandes déviations suivante : pour tout ensemble fermé $A \subset \Omega$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{Q}^\star_{n,x}\left(\frac{1}{n}\alpha_{[nt]} \in A\right) \leq -\inf_{\phi\in A}\left\{ I - \beta F - \inf_{\phi\in\Omega}(I-\beta F)\right\}$$

où $\beta = 1$ dans le cas des listes linéaires et $\beta = 1/2$ dans le cas des queues de priorité et où les fonctionnelles $F : \Omega \longrightarrow [-\infty, +\infty[$ et $I : \Omega \longrightarrow [0, +\infty[$ sont définies par :

$$F(\phi) = \int_0^1 \log(\phi(t))\ dt,$$

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^\star(\dot{\phi}(t))\ dt & si\ \phi \in \mathcal{AC}\ et\ \phi(0) = \phi(1) = 0 \\ +\infty & sinon \end{cases},$$

$\mathcal{AC}$ étant l'ensemble des fonctions absolument continues sur $[0,1]$ et $\Lambda^\star : \mathbb{R} \to [0, +\infty]$ la fonction de taux associée à la marche dynamique

$$\Lambda^\star(x) = \sup_{\lambda\in\mathbb{R}} \left\{ \lambda x - \int_E \log\left(e^\lambda f + e^{-\lambda}(1-f)\right) \mathrm{d}\mu \right\}.$$

La loi du processus de taille $(\alpha_k)_{0\leq k\leq n}$ sous $\mathbb{Q}^\star_{n,x}$ constitue une marche aléatoire dynamique pondérée. Nous utilisons une stratégie analogue à celle du chapitre précédent. L'étude des grandes déviations de la marche dynamique sans pondération est faite, et nous démontrons un théorème du type Mogulskii pour la marche dynamique. Nous utilisons ensuite la méthode de Laplace pour tenir compte de la pondération, mais ici encore, l'irrégularité des fonctionnelles en jeu empêche une application directe et quelques difficultés techniques se présentent. L'étude de la fonctionnelle de taux montre qu'elle admet un unique minimum, et on obtient le résultat suivant :

Proposition 0.2. *Sous les hypothèses du théorème précédent, il existe une fonction $\phi : [0,1] \to [0,1]$ telle que une liste linéaire typique de longueur n :*
– a une taille $\phi(t)n$ au temps $[nt]$, $0 \leq t \leq 1$.
– a un coût de stockage $\left(\int_0^1 \phi(t)\mathrm{d}t\right)\ n^2$.
Les convergences ont lieu avec une rapidité exponentielle.

Nous obtenons des résultats analogues pour les queues de priorité dynamiques et les dictionnaires dynamiques. Le problème variationnel caractérisant la fonction limite est dur à résoudre, nous donnons cependant un exemple de résolution dans le cas d'un système dynamique simple.

Chapitre 1

Quelques éléments de théorie des grandes déviations.

Sommaire

Ce chapitre constitue une introduction rapide à la théorie des grandes déviations. Les principaux outils et méthodes utilisés dans nos travaux y sont présentés. La partie sur le modèle d'Ising, relativement longue, trouve sa place en tant que premier exemple concret d'utilisation des grandes déviations et elle permet également au lecteur de se familiariser avec ce type de modèles en vue de la lecture du chapitre 2.

1.1 La notion de principe de grandes déviations

La théorie des grandes déviations s'intéresse aux événements rares et au calcul asymptotique de leur probabilité dans une échelle exponentielle. Au premier abord, elle peut paraître assez formelle, aussi nous commençons par donner l'idée intuitive de ce qu'est un principe de grandes déviations. Soit X_n une suite de variables aléatoires à valeur dans un espace E. On dit que la suite vérifie un principe de grandes déviations (PGD) de bonne fonction de taux $I : E \to [0, +\infty]$ et de vitesse v_n si "la probabilité que X_n soit proche de $x \in E$ est de l'ordre de $e^{-v_n I(x)}$", et l'on note :

$$\mathbb{P}(X_n \approx x) \asymp e^{-v_n I(x)}.$$

La définition formelle fait intervenir la topologie de E et sera donnée dans le paragrape 1.1.1. Dans les cas que nous considérerons, la vitesse v_n sera toujours égale à n. On parle d'événements rares car si $I(x) \neq 0$, la probabilité $\mathbb{P}(X_n \approx x)$ à laquelle nous nous intéressons décroit très rapidement. Avec cette définition heuristique, il paraît intuitif que les variables X_n se concentrent sur l'ensemble $\{I = 0\}$ où la fonction de taux s'annule. Nous verrons en particulier comment obtenir des résultats de type loi des grands nombres lorsque la fonction de taux s'annule en un seul point.

A titre de premier exemple, considérons le cas où X_n est la moyenne de n variables indépendantes identiquement distribuées de loi gaussienne centrée réduite. La variable X_n suit une loi normale d'espérance 0 et de variance $1/n$, si bien que pour tout intervalle A,

$$\mathbb{P}(\sqrt{n}X_n \in A) = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx.$$

La valeur typique de X_n est de l'ordre de $1/\sqrt{n}$. Soit $\delta > 0$. La probabilité de l'événement $\{|X_n| \geq \delta\}$ tend vers 0, et plus précisément

$$\mathbb{P}(|X_n| \geq \delta) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} e^{-x^2/2} dx,$$

si bien que

$$\frac{1}{n} \log \mathbb{P}(|X_n| \geq \delta) \underset{n \to +\infty}{\to} -\frac{\delta^2}{2}.$$

Ainsi, avec une petite probabilité (de l'ordre de $e^{-n\delta^2/2}$), $|X_n|$ dévie de son comportement typique en prenant des grandes valeurs. Cela explique la terminologie "grandes déviations".

1.1.1 Définitions

Soit E un espace polonais muni de sa métrique $d(.,.)$ et $\mathcal{B}$ la tribu des boréliens sur E, éventuellement complétée. Soit $(\mathbb{P}_n)_{n\in\mathbb{N}}$ une suite de mesures de probabilité sur E.

Definition 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.*

La compacité des ensembles de niveau garantit que sur tout ensemble fermé, une bonne fonction de taux atteint son minimum.

Le PGD s'énonce comme suit.

Definition 1.2. *La suite $(\mathbb{P}_n)_{n\in\mathbb{N}}$ suit un principe de grandes déviations de vitesse v_n et de fonction de taux I si :*
tout fermé $F \in \mathcal{B}$ vérifie la borne supérieure de grandes déviations

$$\limsup_{n\to+\infty} \frac{1}{v_n} \log \mathbb{P}_n(F) \leq - \inf_{x\in F}\{I(x)\},$$

tout ouvert $O \in \mathcal{B}$ vérifie la borne inférieure de grandes déviations

$$\liminf_{n\to+\infty} \frac{1}{v_n} \log \mathbb{P}_n(F) \geq - \inf_{x\in O}\{I(x)\}.$$

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.

En appliquant ces inégalités à l'ensemble E, on voit que l'infimum de I sur E est nul. Notons que la notion de PGD dépend de la structure topologique de l'espace E. Soit $\mathcal{A}$ une base de la topologie de E. La fonction de taux est unique, et caractérisée par la propriété suivante

$$\begin{aligned} I(x) &= \sup\left\{ -\liminf_{n\to+\infty} \frac{1}{v_n} \log \mathbb{P}_n(A) \mid A \in \mathcal{A} \text{ t.q. } x \in A \right\} \\ &= \sup\left\{ -\limsup_{n\to+\infty} \frac{1}{v_n} \log \mathbb{P}_n(A) \mid A \in \mathcal{A} \text{ t.q. } x \in A \right\} \end{aligned}$$

Nous considérerons presque toujours des PGD de vitesse $v_n = n$, et nous omettrons de préciser la vitesse dans ce cas.

1.1.2 Concentration des mesures

Le fait qu'un principe de grandes déviations entraîne un phénomène de concentration de la mesure est pour nous capital. Nous visons dans les applications à déterminer le comportement typique des variables aléatoires représentant certaines quantités d'intérêt. Les résultats que nous obtenons sont de type loi des grands nombres, et se déduisent d'un principe de grandes déviations, en montrant que la bonne fonction de taux atteint son minimum en un unique point. On utilise le résultat suivant, relativement élémentaire, mais essentiel dans nos applications.

Proposition 1.1. *Si $\mathbb{P}_n$ vérifie la borne supérieure de grandes déviations et si la bonne fonction de taux s'annule en un unique point $x^\star$, alors $\mathbb{P}_n$ converge en loi vers la masse de Dirac $\delta_{x^\star}$ avec une vitesse exponentielle.*

Preuve : Soit $(\mathbb{P}_n)_{n\in\mathbb{N}}$ une suite de probabilités vérifiant un PGD de bonne fonction de taux I. Notons $Z = \{I = 0\}$ le fermé où I est nulle, et pour $\varepsilon > 0$, définissons $A_\varepsilon = \{x \in E \mid d(x, Z) \geq \varepsilon\}$. Comme A_ε est fermé, la bonne fonction de taux I y atteint son minimum $I_\varepsilon > 0$. La borne supérieure de grandes déviations s'écrit

$$\limsup_{n\to+\infty} \frac{1}{n} \log \mathbb{P}_n(A_\varepsilon) \leq -I_\varepsilon.$$

On en déduit que pour n suffisamment grand, $\mathbb{P}_n(A_\varepsilon) \leq e^{-nI_\varepsilon/2}$.
Dans le cas où la fonction de taux I s'annule en un unique point $x^\star$, on obtient que pour n suffisamment grand,

$$\mathbb{P}_n\left(\{x \in E \mid d(x, x^\star) \leq \varepsilon\}\right) \geq 1 - e^{-nI_\varepsilon/2}.$$

□

Si les mesures $\mathbb{P}_n$ sont les mesures images de variables aléatoires X_n définies sur un même espace probabilisé $(\Omega, \mathbb{P})$, on a pour n grand

$$\mathbb{P}(d(X_n, x^\star) \geq \varepsilon) \leq e^{-nI_\varepsilon/2}.$$

On en déduit la loi faible des grands nombres : X_n converge en probabilité vers $x^\star$. Comme pour tout $\alpha > 0$, la série $\sum_{n\geq 0} e^{-n\alpha}$ converge, le lemme de Borel-Cantelli permet de prouver la loi forte des grands nombres : X_n converge presque sûrement vers $x^\star$.

Remarque 1.1. *La notion de concentration de mesures est assez générale et dépasse le cadre des grandes déviations. Dans certaines situations, les estimées précises de grandes déviations ne sont pas disponibles, mais l'on dispose d'inégalités de concentration. Nous pensons en particulier aux inégalités de concentration pour les martingales à différence bornés, et aux inégalités de concentration de Talagrand [51, 52].*

1.2 Quelques outils généraux

Une des grandes forces de la théorie des grandes déviations est que plusieurs outils permettent de "transférer" un PGD d'une situation à une autre. Ainsi, on commence souvent par montrer un PGD dans une situation simple, ou par utiliser un PGD connu, puis on applique une transformation pour transférer le PGD dont on dispose à la situation qui nous intéresse. Nous présentons ici trois résultats : l'utilisation d'approximations exponentielles, le principe de contraction et la méthode de Laplace. Ces techniques sont utilisées par la suite. La méthode de Laplace est au cœur de tous les résultats que nous développons dans les applications, et c'est elle qui permet d'écrire la fonction de taux dont on cherche ensuite les minimas.

1.2.1 Approximations exponentielles

L'idée intuitive est la suivante : deux suites de mesures de probabilité $(\mathbb{P}_n)_{n\in\mathbb{N}}$ et $(\mathbb{Q}_n)_{n\in\mathbb{N}}$ suffisamment proches ont les mêmes propriétés du point de vue des grandes déviations. La bonne notion est celle de mesures exponentiellement équivalentes :

Définition 1.1. *Soient (E,d) un espace métrique, et des mesures de probabilité $(\mathbb{P}_n)_{n\in\mathbb{N}}$ et $(\mathbb{Q}_n)_{n\in\mathbb{N}}$ sur E. On dit que les mesures $(\mathbb{P}_n)_{n\in\mathbb{N}}$ et $(\mathbb{Q}_n)_{n\in\mathbb{N}}$ sont exponentiellement équivalentes si il existe des espaces probabilisés $(\Omega,\mathcal{B}_n,\mu_n)_{n\in\mathbb{N}}$ et deux familles de variables aléatoires à valeurs dans E, $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$, de lois jointes $(\mu_n)_{n\in\mathbb{N}}$ et marginales $(\mathbb{P}_n)_{n\in\mathbb{N}}$ et $(\mathbb{Q}_n)_{n\in\mathbb{N}}$ respectivement telles que pour tout $\delta>0$, l'ensemble $\{\omega\in\Omega \mid d(X_n,Y_n)>\delta\}$ est mesurable et*

$$\limsup_{n\to+\infty}\frac{1}{n}\log\mu_n(\ d(X_n,Y_n)>\delta\ )=-\infty.$$

Des mesures de probabilité exponentiellement équivalentes ont les mêmes propriétés du point de vue des grandes déviations, comme l'énonce le théorème suivant ([19] section 4.2.2) :

Théorème 1.1. *Si les mesures $(\mathbb{P}_n)_{n\in\mathbb{N}}$ vérifient un PGD de bonne fonction de taux I et sont exponentiellement équivalentes aux mesures $(\mathbb{Q}_n)_{n\in\mathbb{N}}$, alors le même PGD est vérifié par $(\mathbb{Q}_n)_{n\in\mathbb{N}}$.*

1.2.2 Principe de contraction

Le principe de contraction explique comment une application continue permet de transférer un PGD d'un espace à un autre. De démonstration élémentaire (cf [19] section 4.2), ce théorème est néanmoins important :

Théorème 1.2. *Soient E et F deux espaces de Hausdorff et $f : E \to F$ une fonction continue. Soient $(\mathbb{P}_n)_{n\in\mathbb{N}}$ des mesures de probabilité sur E satisfaisant un PGD de bonne fonction de taux I. Alors la fonctionnelle $J : F \mapsto [0, +\infty]$ définie par*

$$J(y) = \inf\{I(x) \mid x \in E, y = f(x)\}$$

est une bonne fonction de taux et la suite des mesures images $(\mathbb{P}_n \circ f^{-1})_{n\in\mathbb{N}}$ vérifie un PGD contrôlé par J.

Des extensions sont disponibles pour traiter le cas de fonctions presque continues (approximations exponentielles de fonction continue).

1.2.3 Méthode de Laplace

La méthode de Laplace permet d'estimer asymptotiquement certaines intégrales exponentielles. C'est un outil très utile dans les applications. Par exemple, en mécanique statistique, elle permet souvent de déterminer le comportement asymptotique de la fonction de partition. Cette méthode est au cœur de nos travaux, aussi nous allons la présenter avec de plus amples détails, les preuves sont détaillées dans [19] section 4.3.

Théorème 1.3. *Supposons que les mesures de probabilité $(\mathbb{P}_n)_{n\in\mathbb{N}}$ sur E vérifient un PGD de bonne fonction de taux $I : E \to [0, +\infty]$. Soit $\theta : E \to \mathbb{R}$ une fonction continue bornée. Posons,*

$$Z_n = \int_E e^{n\theta(x)} d\mathbb{P}_n(x).$$

Alors,

$$\lim_{n\to+\infty} \frac{1}{n} \log Z_n = -\inf_{x\in E}\{I(x) - \theta(x)\}, \tag{1.1}$$

et la suite de probabilités $(\mathbb{Q}_n)_{n\in\mathbb{N}}$ définies par

$$\frac{d\mathbb{Q}_n}{d\mathbb{P}_n}(x) = \frac{1}{Z_n} e^{n\theta(x)},$$

vérifie un PGD de bonne fonction de taux

$$J(x) = I(x) - \theta(x) - \inf_{y\in E}\{I(y) - \theta(y)\}.$$

Remarque : L'hypothèse de bornitude pour θ garantit que la fonctionnelle J soit une bonne fonction de taux. Elle peut être relachée si on se contente de l'asymptotique de Z_n : la condition de queue

$$\lim_{M\to+\infty} \limsup_{n\to+\infty} \frac{1}{n} \log \int_E e^{n\theta(x)} \mathbf{1}_{\{\theta(x)\geq M\}} d\mathbb{P}_n(x) = -\infty, \tag{1.2}$$

ou la condition de moment

$$\limsup_{n\to+\infty} \frac{1}{n} \log \int_E e^{\gamma n\theta(x)} d\mathbb{P}_n(x) <= \infty, \quad \text{pour un } \gamma > 1$$

suffisent à garantir l'équation (1.1).

La preuve de ce théorème repose sur les deux lemmes suivants :

Lemme 1.1. *Si $\theta : E \to [-\infty, +\infty)$ est semi continue supérieurement et vérifie la condition de queue (1.2), et si la borne supérieure de grandes déviations est satisfaite avec la bonne fonction de taux I, alors*

$$\limsup_{n\to+\infty} \frac{1}{n} \log Z_n \leq - \inf_{x\in E} \{I(x) - \theta(x)\}.$$

Lemme 1.2. *Si $\theta : E \to (-\infty, +\infty]$ est semi continue inférieurement et si la borne inférieure de grandes déviations est satisfaite avec la bonne fonction de taux I, alors*

$$\liminf_{n\to+\infty} \frac{1}{n} \log Z_n \geq - \inf_{x\in E} \{I(x) - \theta(x)\}.$$

Dans le chapitre 2, nous utilisons ce théorème avec l'hypothèse θ continue bornée. Nous obtenons donc un PGD complet avec une bonne fonction de taux dont nous étudions les minimas pour obtenir une loi des grands nombres. Dans les chapitres 3 et 4, nous voulons appliquer la méthode de Laplace avec une fonctionnelle θ seulement semi-continue supérieurement, nous utilisons alors le lemme 1.1.

1.3 Principes de grandes déviations dans $\mathbb{R}^d$

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

$$S_n = \sum_{k=1}^{n} X_k$$

où $(X_k)_{k\geq 1}$ est une suite de variables aléatoires indépendantes et identiquement distribuées à valeurs réelles. Supposons que les X_k sont centrées et que pour tout $t \in \mathbb{R}$, $\exp(tX_1)$ est intégrable. On note Λ la fonction génératrice des cumulants de X_1 donnée par

$$\Lambda(t) = \log \mathbb{E}(\exp(tX_1)),$$

et $\Lambda^\star$ sa transformée de Fenchel-Legendre

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

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

$$\mathbb{P}(S_n - nx \geq 0) \leq \mathbb{E}(e^{t(S_n - nx)}) = e^{-n(tx - \Lambda(t))}.$$

En optimisant sur le paramètre t, on obtient

$$\frac{1}{n} \log \mathbb{P}(S_n/n \geq x) \leq - \sup_{t \geq 0} \{tx - \Lambda(t)\}.$$

Comme les variables X_k sont centrées, $\sup_{t \geq 0}\{tx - \Lambda(t)\} = \Lambda^\star(x)$ et

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

Cette inégalité est une borne supérieure de grandes déviations. On montre que la suite $(S_n/n)_{n \geq 1}$ satisfait un PGD de bonne fonction de taux $\Lambda^\star$ et de vitesse n. Ce résultat constitue le *théorème de Cramer*. Nous présentons ci-après une généralisation de ce théorème, qui permet d'obtenir un PGD à partir du comportement asymptotique des fonctions génératrices des cumulants.

1.3.1 Le théorème de Gärtner-Ellis

Avant d'énoncer le théorème de Gärtner-Ellis, rappelons la définition d'une fonction essentiellement lisse.

Définition 1.2. *Soit $I : \mathbb{R}^d \to]-\infty\,;\,+\infty]$ une fonction convexe de domaine $\mathcal{D} = \{x\,;\, I(x) < +\infty\}$. 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 $\mathcal{D}$,
(iii) si pour toute suite (x_n) dans l'intérieur de $\mathcal{D}$ convergeant vers un point de la frontière de cet intérieur, $I'(x_n)$ converge en norme vers $+\infty$.

Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$. On munit $\mathbb{R}^d$ du produit scalaire canonique $< .,. >$. Notons

$$\Lambda_n(t) = \log \mathbb{E}(e^{<t, X_n>}), \quad t \in \mathbb{R}^d,$$

la suite des fonctions génératrices des cumulants. Supposons que, pour tout $t \in \mathbb{R}^d$, $\Lambda_n(v_n t)/v_n$ converge vers $\Lambda(t)$ (éventuellement égal à $+\infty$) et que 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

$$\Lambda^\star(x) = \sup_{t \in \mathbb{R}^d} \{< t, x > - \Lambda(t)\}.$$

Dans ce cas, le théorème de Gärtner-Ellis fournit un PGD :

Théorème 1.4. *Sous ces hypothèses, la suite de variables aléatoires $(X_n)_{n\geq 1}$ suit un principe de grandes déviations de vitesse v_n 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) et la borne supérieure pour ces nouveaux processus. Enfin l'analyse convexe est fortement utilisée pour étudier les propriétés de la fonctionnelle $\Lambda^\star$. Pour une démonstration précise, voir par exemple [19], section 2.3.

1.3.2 Le modèle d'Ising en dimension 1

Nous faisons ici une digression dans cet exposé général sur les bases de la théorie des grandes déviations pour présenter le modèle d'Ising. Nous avons plusieurs raisons. C'est l'ocasion d'introduire un modèle simple de mécanique statistique, et d'y voir fonctionner la théorie des grandes déviations (on applique en particulier le théorème de Gärtner-Ellis). De plus, cela nous permet de nous familiariser avec ce type de modèles qui seront réintroduits dans le chapitre 2 pour modéliser la dénaturation de l'ADN.

Présentation du modèle.
Il fut introduit pour décrire les métaux ferromagnétiques. On imagine que sur un cercle sont placés régulièrement N atomes étiquetés de 1 à N. Chaque atome possède un moment magnétique ou spin σ qui peut être orienté vers le haut ou vers le bas. Le spin de l'atome situé à l'abscisse i est noté σ_i et vaut ± 1 selon son orientation (par convention, $+1$ correspond à un spin orienté vers le haut). Le système est caractérisé par la donnée des N spins et l'on note $\sigma = (\sigma_i)_{i=1}^N \in \{+1, -1\}^N$ une configuration. On appelle $\Omega_N = \{+1, -1\}^N$ l'espace des configurations.

A chaque configuration σ est associée son énergie ou hamiltonien $H_N(\sigma)$. Cet hamiltonien prend en compte :
- l'interaction du spin σ_i avec un champ magnétique extérieur supposé homogène caractérisé par son intensité $b \in \mathbb{R}$. Cette interaction apporte une contribution $b\sigma_i$ au hamiltonien.
- une interaction entre les spins σ_i et σ_j, pour $i \neq j$. On se place dans le cadre d'une interaction entre plus proches voisins : il y a interaction entre σ_i et σ_j seulement pour $|j - i| = 1$. Cette interaction apporte une contribution $-a\sigma_i\sigma_{i+1}$. On suppose la constante $a > 0$, cela signifie que les spins voisins ont tendance à s'aligner.

L'hamiltonien est donné par :

$$H_N(\sigma) = b\sum_{i=1}^{N}\sigma_i - a\sum_{i=1}^{N}\sigma_i\sigma_{i+1},$$

où on utilise la convention $\sigma_{N+1} = \sigma_1$ (condition aux bords périodique). On définit pour chaque configuration deux quantités macroscopiques, l'aimantation $M_N(\sigma)$ et le périmètre $R_N(\sigma)$, par

$$M_N(\sigma) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i,$$

$$R_N(\sigma) = \frac{1}{N}\sum_{i=1}^{N}\frac{1-\sigma_i\sigma_{i+1}}{2}.$$

L'hamiltonien s'écrit en fonction de ces quantités :

$$H_N(\sigma) = N[bM_N(\sigma) - a + 2aR_N(\sigma)].$$

D'un point de vue physique, l'aimantation et le périmètre sont des grandeurs accessibles à la mesure, contrairement à la configuration exacte σ ou bien à la valeur particulière d'un spin σ_i. On appelle ces quantités les observables. Elles caractérisent l'état macroscopique du système. L'état du système lui n'est pas totalement déterminé, les quantités microscopiques (ici les spins) fluctuent très rapidement. On quantifie la probabilité de trouver le système dans telle ou telle configuration. L'état du système est caractérisé par une mesure de probabilité sur l'espace des configurations Ω_N, appelée mesure de Gibbs.

La mesure de Gibbs correspondant à l'hamiltonien H_N et à la température $T = \frac{1}{\beta}$ est définie par

$$\pi_N^\beta(\sigma) = \frac{1}{Z_N^\beta}e^{-\beta H_N(\sigma)}$$

où $Z_N^\beta := Z_N^\beta(a,b)$ est la fonction de partition

$$Z_N^\beta = \sum_{\sigma\in\Omega_N} e^{-\beta H_N(\sigma)}.$$

Sous cette mesure, les configurations d'énergie minimale sont les plus probables. Lorsque la température diminue (β augmente), la mesure se concentre. A température infinie ($\beta = 0$), la mesure de Gibbs est la mesure uniforme sur Ω_N. A température nulle ($\beta = \infty$), la mesure de Gibbs est la mesure uniforme sur l'ensemble des configurations d'énergie minimale.

Grandeurs moyennes.
Il s'agit de relier les valeurs moyennes de l'aimantation, du périmètre, de l'hamiltonien à la fonction de partition Z_N^β. Si G est une grandeur définie sur Ω_N,

la moyenne de G pour la mesure π_N^β est

$$< G >_{\pi_N^\beta} = \sum_{\sigma \in \Omega_N} G(\sigma)\pi_N^\beta(\sigma) = \frac{1}{Z_N^\beta} \sum_{\sigma \in \Omega_N} G(\sigma) e^{-\beta H_N(\sigma)}.$$

On définit l'énergie libre par site

$$F_N^\beta = \frac{1}{N} \log(Z_N^\beta).$$

Elle permet d'exprimer simplement les valeurs moyennes de l'aimantation, du périmètre et de l'hamiltonien pour $\beta > 0$:

$$< M_N >_{\pi_N^\beta} = -\frac{1}{\beta} \frac{\partial}{\partial b} F_N^\beta,$$

$$< R_N >_{\pi_N^\beta} = \frac{1}{2\beta} \frac{\partial}{\partial a} F_N^\beta,$$

$$< \frac{1}{N} H_N >_{\pi_N^\beta} = -\frac{\partial}{\partial \beta} F_N^\beta.$$

Calcul de la fonction de partition.
Le principe de ce calcul est basé sur la méthode de matrice de transfert. Il s'agit d'interpréter le calcul de la fonction de partition en termes d'opérations sur des matrices.

La fonction de partition s'écrit

$$Z_N^\beta = \sum_{\sigma_1, \cdots, \sigma_N \in \{+1,-1\}} \prod_{i=1}^{N} e^{-\beta b \sigma_i + \beta a \sigma_i \sigma_{i+1}}.$$

On introduit la matrice 2×2, dite matrice de transfert,

$$L = \begin{pmatrix} e^{-\beta b + \beta a} & e^{-\beta b - \beta a} \\ e^{\beta b - \beta a} & e^{\beta b + \beta a} \end{pmatrix}.$$

Alors la fonction de partition s'écrit simplement

$$Z_N^\beta = Tr(L^N).$$

Notons λ_1 et λ_2 les valeurs propres de L, avec $\lambda_1 \geq \lambda_2$, on a les formules

$$Z_N^\beta = \lambda_1^N + \lambda_2^N,$$

$$\lambda_1 = e^{\beta a} \left(\cosh(\beta b) + \sqrt{\cosh^2(\beta b) - 1 + e^{-4\beta a}} \right),$$

$$\lambda_2 = e^{\beta a} \left(\cosh(\beta b) - \sqrt{\cosh^2(\beta b) - 1 + e^{-4\beta a}} \right).$$

A partir de cela, les valeurs moyennes de l'aimantation, du perimètre, de l'hamiltonien ... sont calculables explicitement, à partir de l'énergie libre par site

$$F_N^\beta = \log(\lambda_1) + \frac{1}{N}\log\left(1 + (\frac{\lambda_1}{\lambda_2})^N\right).$$

Lorsque $N \to \infty$, F_N et ses dérivées partielles convergent uniformément sur les compacts en $\beta > 0, a, b$. On en déduit que les valeurs moyennes de l'aimantation, du périmètre ont une limite lorsque $N \to \infty$. On note $\rho(L) = \lambda_1$ le rayon spectral de L. On a :

$$F_\infty^\beta = \lim_{N\to\infty} F_N^\beta = \log(\rho(L))$$

$$M_\infty^\beta = \lim_{n\to\infty} < M_N >_{\pi_N^\beta} = -\frac{1}{\beta}\frac{\partial}{\partial b}F_\infty^\beta$$

$$R_\infty^\beta = \lim_{N\to\infty} < R_N >_{\pi_N^\beta} = \frac{1}{2\beta}\frac{\partial}{\partial a}F_\infty^\beta.$$

Grandes déviations pour (M_N, R_N)**.**
Le théorème de Gärtner-Ellis permet d'obtenir un PGD pour la loi des deux observables M_N et R_N sous la mesure de Gibbs π_N^β.

Soit $t_1, t_2 \in \mathbb{R}$. On commence par calculer les moments exponentiels :

$$< e^{N(t_1 M_N + t_2 R_N)} >_{\pi_N^\beta} = \sum_{\sigma\in\Omega_N} e^{N(t_1 M_N + t_2 R_N)}\pi_N^\beta(\sigma) = \frac{Z_N^\beta(b - \frac{t_1}{\beta}, a + \frac{t_2}{\beta})}{Z_N^\beta(b, a)}$$

On a donc

$$\frac{1}{N}\log < e^{N(t_1 M_N + t_2 R_N)} >_{\pi_N^\beta} = F_N^\beta(b - \frac{t_1}{\beta}, a + \frac{t_2}{\beta}) - F_N^\beta(b, a)$$

Cette quantité a une limite lorsque $N \to \infty$ qui vaut :

$$\Lambda^\beta(t_1, t_2) = F_\infty^\beta(b - \frac{t_1}{\beta}, a + \frac{t_2}{\beta}) - F_\infty^\beta(b, a).$$

D'après le théorème de Gärtner-Ellis, la loi de (M_N, R_N) sous la mesure π_N^β vérifie un principe de grandes déviations de vitesse N et de bonne fonction de taux :

$$I^\beta(m, r) = \sup\left\{\Lambda^\beta(t_1, t_2) - mt_1 - rt_2 \ ; \ (t_1, t_2) \in \mathbb{R}^2\right\}.$$

1.4 Principes de grandes déviations fonctionnels

La théorie des grandes déviations admet de nombreux développements concernant des mesures sur des espaces de dimension infinie. Citons par exemple le théorème de Sanov donnant un PGD pour la mesure empirique d'un échantillon de variables aléatoires identiquement distribuées, ou encore le théorème de Schilder donnant un PGD pour des mesures liées au mouvement brownien. Ces résultats théoriques ont des applications importantes : le théorème de Sanov est à la base de nombreux résultats portant sur les systèmes de particules en interaction, le théorème de Schilder est un outil important dans la théorie des équations différentielles stochastiques. Dans nos travaux, nous utilisons le théorème de Mogulskii donnant un principe de grandes déviations fonctionnel pour les trajectoires des marches aléatoires.

1.4.1 Le théorème de Mogulskii

Soit $X_1, X_2, \cdots$ une suite de variables aléatoires i.i.d. à valeurs dans $\mathbb{R}^d$, telles que pour tout $t \in \mathbb{R}^d$,

$$\Lambda(t) = \log \mathbb{E}(e^{<t,X_1>}) < \infty.$$

Le théorème de Cramer affirme que la moyenne empirique

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

vérifie un PGD de bonne fonction de taux $\Lambda^\star$ définie par

$$\Lambda^\star(x) = \sup_{t \in \mathbb{R}^d} \{< t, x > -\Lambda(t)\}.$$

Le théorème de Mogulskii établit un PGD pour le processus $S_n(.)$ défini par

$$S_n(t) = \frac{1}{n} \sum_{i=1}^{[nt]} X_i \ , \quad 0 \leq t \leq 1,$$

dans l'espace Ω des fonctions càd-làg sur $[0,1]$ muni de la norme $\|.\|_\infty$. Notons $\widetilde{\mathbb{P}}_n$ sa loi.

Théorème 1.5. *La suite de mesures $(\widetilde{\mathbb{P}}_n)_{n\in\mathbb{N}}$ vérifie un PGD de bonne fonction de taux $I : \Omega \to [0, +\infty]$ définie par*

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^\star(\dot{\phi}(t))dt & \text{si } \phi \in \mathcal{AC} \text{ et } \phi(0) = 0 \\ +\infty & \text{sinon} \end{cases},$$

où $\mathcal{AC}$ est le sous-espace de Ω constitué des fonctions absolument continues.

La preuve de ce théorème repose sur un PGD pour les vecteurs aléatoires $(S_n(t_1), \cdots, S_n(t_k))$, $0 < t_1 < \cdots < t_k \leq 1$. Le théorème de Dawson-Gärtner (cf [19]) permet d'en déduire un principe de grandes déviations pour le processus $S_n(.)$ dans l'ensemble des fonctions de $[0,1]$ dans $\mathbb{R}^d$ muni de la topologie produit, en considérant cet espace comme la limite projective d'espaces de dimension finie où des PGD sont disponibles. L'utilisation d'une approximation exponentielle continue du processus permet de passer au PGD sur l'espace Ω muni de la topologie de la norme uniforme.

1.4.2 Cas de la marche aléatoire dynamique

La marche aléatoire dynamique introduite par N.Guillotin-Plantard dans [29, 30, 31] permet de travailler avec des marches aléatoires à accroissements non stationnaires. La loi des incréments fait intervenir un système dynamique qui permet d'introduire une inhomogénéité temporelle dans le processus. Dans le chapitre 4, les propriétés de grandes déviations de la marche aléatoire dynamique sont étudiées dans le cadre d'une application à l'informatique théorique. Nous rappelons ici les résultats de grandes déviations obtenus.

Soit $S = (E, \mathcal{A}, \mu, T)$ un système dynamique où $(E, \mathcal{A}, \mu)$ est un espace probabilisé et T une transformation de E préservant la mesure μ. Soit $d \geq 1$ et $f_1, \cdots, f_d$ des fonctions définies sur E à valeurs dans $[0, \frac{1}{d}]$. Soit $(X_i)_{i\geq 1}$ une suite de vecteurs indépendants à valeurs dans $\mathbb{Z}^d$. Soit $x \in E$ un point fixé et $(e_j)_{i\leq j\leq d}$ la base canonique de $\mathbb{Z}^d$. Pour tout $i \geq 1$, la loi du vecteur aléatoire X_i est donné par

$$\mathbb{P}_x(X_i = z) = \begin{cases} f_j(T^i x) & \text{si } z = e_j \\ \frac{1}{d} - f_j(T^i x) & \text{si } z = -e_j \\ 0 & \text{sinon.} \end{cases}$$

La marche aléatoire dynamique est définie par $S_0 = 0$ et pour $n \geq 1$,

$$S_n = \sum_{i=1}^{n} X_i.$$

Les propriétés de grandes déviations de la marche dynamique sont données dans les deux théorèmes suivants qui seront démontrés au chapitre 4.

Théorème 1.6. *Pour μ-presque tout point $x \in E$, la suite $(\frac{1}{n}S_n)_{n\geq 1}$ vérifie dans $\mathbb{R}^d$ un PGD de bonne fonction de taux*

$$\Lambda^\star(x) = \sup_{t\in\mathbb{R}^d}\{< t, x > -\Lambda(t)\}$$

avec

$$\Lambda(t) = \mathbb{E}\left(\log(\sum_{j=1}^{d} e^{t_j} f_j + (\frac{1}{d} - f_j)e^{-t_j})|\mathcal{I}\right),$$

$\mathcal{I}$ étant la tribu engendrée par les points fixes de T.

Définissons le processus $S_n(.)$ par $S_n(t) = \frac{1}{n}\sum_{i=1}^{[nt]} X_i,\ 0 \leq t \leq 1$.

Théorème 1.7. *Supposons le système dynamique $(E, \mathcal{A}, \mu, T)$ uniquement ergodique et les fonctions $f_1, \cdots, f_d$ continues. Alors pour tout $x \in E$, le processus $S_n(.)$ vérifie dans Ω un PGD de bonne fonction de taux*

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^\star(\dot{\phi}(t))dt & \text{si } \phi \in \mathcal{AC} \text{ et } \phi(0) = 0 \\ +\infty & \text{sinon} \end{cases} .$$

Chapitre 2

A stochastic model for DNA denaturation.

Sommaire

Ce chapitre est tiré de l'article [21] publié dans *Journal of Statistical Physics.*

Abstract

We consider Benham's model for strand separation in supercoiled circular DNA. This is a mean field model in external inhomogeneous field, conditioned to small values of the perimeter. Under some conditions on the external field, we prove a large deviations principle for the distribution of the magnetization under the Gibbs measure. The rate function strongly depends on the structure of the external field. It allows us to prove a law of large numbers and to study denaturation as a function of the temperature and the superhelical density.

2.1 Introduction

Fundamental biological mechanisms such as replication and transcription of DNA require the two strands of the DNA double helix to separate. The separation of the two strands - called denaturation - can be partial or total. Denaturation depends on several factors such as the temperature, the composition of the DNA sequence and the geometry of the DNA polymer. Benham proposes a mathematical model for the process of denaturation ([2, 3, 4]). His model is based on statistical mechanics ideas and takes into account the temperature, the nature of the bases and the superhelicity of DNA. He also develops algorithms to locate the regions where the denaturation is highly susceptible to occur. In [45], Mazza computes the thermodynamic limit of Benham's model in the homopolymer case, and for some special types of heteropolymer. Denaturation only occurs in a few regions, and then eventually expands. Therefore Mazza proposes a modification of Benham's model focusing on configurations having a small number of "denaturation bubbles". He computes the thermodynamic limit of this new model for some types of DNA sequence and shows that the model exhibits phenomena of phase transition. The present work is motivated by extending these computations to more general heteropolymers. Identifying the cases where it is possible is a large part of the work, and is connected with the notion of the structure of the DNA sequence.

We introduce Benham's model in a somewhat formal way, that is in the framework of one-dimensional spin model in inhomogeneous external field. Consider a circular graph with N sites, labeled successively $i = 1, \cdots, N$. This graph stands for a circular DNA heteropolymer of length N. This polymer is a double helix, each strand of the helix consists in N nucleotides. Each vertex represents a pair of nucleotides. At each site there is a spin denoted by σ_i taking values

$\{-1,+1\}$. For convenience, we define the variables

$$n_i = \frac{1+\sigma_i}{2} = \begin{cases} +1 & \text{if} \quad \sigma_i = +1 \\ 0 & \text{if} \quad \sigma_i = -1 \end{cases}, \quad i = 1, \cdots, N.$$

The meaning of $n_i = 0$ is that the bases of the double helix at site i are linked by an hydrogen bond, the link is closed, and $n_i = 1$ means that this bond is broken, the link is open. Let $\Omega_N = \{-1,+1\}^N$ be the configuration space. We define some macroscopic quantities on the configuration space. The magnetization of a configuration $\sigma \in \Omega_N$ is defined by

$$M_N(\sigma) = \frac{1}{N}\sum_{i=1}^{N} n_i \in [0,1].$$

The magnetization stands for the proportion of denatured bonds and measures the denaturation : $M_N = 0$ means that all links are closed, the DNA polymer is not denatured, $M_N = 1$ means that all links are open, the polymer is totally denatured. The field B_N is given by a sequence of reals $(b_i^N)_{1\leq i\leq N}$. The value of the field at site i represents the energy of the bond between the bases located at site i. As a nucleotide pair is either $A+T$ or $G+C$, the field takes only two values denoted by b_{AT} and b_{GC}. The AT-links are formed of 2 hydrogen bonds and the GC-links consist in 3 hydrogen bonds so that $b_{GC} > b_{AT}$. The interaction of the spin system with the external field is measured by the localized magnetization

$$M_{NB_N}(\sigma) = \frac{1}{N}\sum_{i=1}^{N} b_i^N n_i.$$

It is a measure of the energy needed to break all the bonds and get the configuration σ. The perimeter of a configuration $\sigma \in \Omega_N$ is defined by

$$R_N(\sigma) = \frac{1}{2}\sum_{i=1}^{N} \mathbb{1}_{\sigma_i \neq \sigma_{i+1}}.$$

We make use of the circular boundary conditions : the site N is regarded as being followed by site 1, so that σ_{N+1} has to be seen as σ_1. This comes from the circular structure of the graph and ensures that the system is invariant under translation. Because of the periodic boundary conditions, the perimeter is an integer. It represents the number of connected domains of denatured bonds or "denaturation bubbles".

The Hamiltonian introduced by Benham is

$$H_N(\sigma) = aR_N(\sigma) + NF(M_N(\sigma), M_{NB_N}(\sigma)) \tag{2.1}$$

where F is the function defined by

$$F(m, \tilde{m}) = \frac{2\pi^2 C K_0}{4\pi^2 C + K_0 m}(\kappa + \frac{m}{A})^2 + \tilde{m}.$$

In this formula, κ is the superhelicity of the DNA polymer and is considered as a parameter and the other terms are biological constants. Sun, Mezei, Fye and Benham [50], the following values are given : at 0.01 molar Na^+ concentration and temperature $T = 310$ K, $a = 10.5$ kcal/mol, $b_{AT} = 0.258$ kcal/mol, $b_{GC} = 1.305$ kcal/mol, $C = 3.6$ kcal/rad^2, $A = 10.4$, $K_0 = 2350\ RT$ with $R = 8.3146$ J/K/mol. For a discussion about this Hamiltonian, the reader should refer to the original works of Benham [2, 3, 4] or to the book by Clote and Backofen [14]. In this Hamiltonian, we take into account the energy of nucleation initiation (through the constant a), the energy of AT or GC separation (through the constants b_{AT} and b_{GC}), the torsional or rotational free energy (through the constant C) and the free energy associated to the residual linking number (through the constant K_0).

Let ρ_N be the uniform probability measure on Ω_N. The Gibbs measure is defined by

$$\pi_N(\sigma) = \frac{1}{Z_N} e^{-H_N(\sigma)} \rho_N(\sigma)$$

where the normalization factor Z_N is the partition function defined by

$$Z_N = \int_{\Omega_N} e^{-H_N(\sigma)} \rho_N(d\sigma).$$

What we are interested in is the asymptotic behavior of the Gibbs measure when it is conditioned on untypical small values of the perimeter. For $r_N \in \mathbf{N}$, define :

$$\rho_{N,r_N} = \rho_N(\ .\ \mid R_N \leq r_N),$$

$$\pi_{Nr_N} = \pi_N(\ .\ \mid R_N \leq r_N).$$

We suppose that the typical equilibrium state of the DNA complex is distributed according to π_{Nr_N}. In [45], Mazza shows that the small perimeter assumption ensures the possibility of phase transitions, that is the possibility for the existence of a stable and robust denatured state when the superhelical density is small enough. Conditioning on small values of the perimeter makes the Ising nearest-neighbour structure to disappear in the limit, at least for the aspects considered in this paper. So, we can consider the small perimeter condition as a slight modification of Benham's model that allows to compute the thermodynamic limit effectively. The justification of this assumption relies on the principle of equivalence of ensembles (see J.T. Lewis and al. [38] and references herein). Roughly speaking, this principle states that in the thermodynamic limit, the microcanonical measures and the grand canonical measures

are equivalent. Hence we believe that the thermodynamic limit of the conditioned measures is equivalent to the thermodynamic limit of the Benham's measures with the value a equal to infinity. This is relevant since the biological constants satisfy $a \gg b_{AT}, b_{GC}$. In this paper, we present an extension of Mazza's results for more general heteropolymer.

Question : What is the asymptotic behavior of the magnetization M_N and the localized magnetization M_{NB_N} under the conditioned Gibbs measure π_{Nr_N} when $r_N << N$?
In order to observe an asymptotic behavior for the localized magnetization M_{NB_N}, we have to impose that the external fields B_N converge in some sense. The most interesting feature of this study is that the influence of the field on the magnetizations M_N and M_{NB_N} is explicited. We obtain indeed a law of large numbers for the pair (M_N, M_{NB_N}), i.e. the pair converges to a deterministic limit denoted by $(M_\infty, \widetilde{M}_\infty)$. This limit can be evaluated and strongly depends on the external fields B_N. The term M_∞ represents the limit proportion of broken links, i.e. the limit denaturation. The term $\widetilde{M}_\infty$ stands for the amount of energy needed to break all the bounds, it gives information on the localization of denaturation. In order to prove the law of large numbers, we study the large deviations properties of the magnetization (M_N, M_{NB_N}).We firstly prove a large deviations principle (LDP) for the pair, and then deduce the law of large numbers.

Our paper is organized as follows. The next section is devoted to the exposition of the results : we state and discuss the hypotheses, give some examples, state the LDP for the distribution of $(M_N, M_{NB_N}, R_N/N)$ under the conditioned Gibbs measure π_{Nr_N}, and deduce a law of large numbers. In the third section, the results are proved. The last section is devoted to applications : we study DNA denaturation as a function of the superhelical density κ, and give numerical computations (see Figure 1 and 2).

2.2 Statement of the results

2.2.1 Hypotheses and examples

In the sequel, for $N \geq 1$, let $B_N = (b_i^N)_{1 \leq i \leq N} \in \{b_{AT}, b_{GC}\}^N$ be external fields and $r_N \geq 1$ be integers. We suppose that the sequence of integers $r_N \geq 1$ satisfies the small perimeter assumption :

- $(\mathbf{H_0})$: $\lim_{N \to +\infty} \frac{r_N}{N} = 0.$

Hypotheses on the sequence of external fields

Let $(T_N)_{N\geq 1}$ be a sequence of integers. We define the average of the field B_N with scale T_N and we denote by $\overline{B}_N = (\bar{b}_i^N)_{1\leq i\leq N} \in [b_{AT}, b_{GC}]^N$ the field defined by

$$\bar{b}_i^N = \frac{1}{T_N} \sum_{k=i}^{i+T_N-1} b_k^N \ , \ 1 \leq i \leq N \tag{2.2}$$

The periodic boundary condition $b_{N+i}^N = b_i^N$ is used.

Roughly speaking, the first hypothesis (H_1) is that the averaged field $\overline{B}_N$ is almost constant with at most r_N values. Denote by L_{Nr_N} the set of fields $G \in \mathbb{R}^N$ which are constant on r_N subintervals : that is $G = (g_i)_{1\leq i\leq N}$ belongs to L_{Nr_N} if and only if there exist integers $0 = l_0 < l_1 \cdots < l_{r_N} = N$ such that g_i is constant on each of the r_N sets of the form $l_k < i \leq l_{k+1}$.
The first hypothesis is :

• $(\mathbf{H_1})$: For each $N \geq 1$, there exists a field $\widetilde{B}_N = (\tilde{b}_i^N)_{1\leq i\leq N} \in L_{Nr_N}$ such that the distance $\delta_N = \frac{1}{N}\sum_{i=1}^N |\bar{b}_i^N - \tilde{b}_i^N|$ has limit zero as $N \to +\infty$.

Since $\bar{b}_i^N \in [b_{AT}, b_{GC}]$, we can also suppose that $\tilde{b}_i^N \in [b_{AT}, b_{GC}]$.

For each field $B_N = (b_i^N) \in \mathbb{R}^N$, we denote by $\mu(B_N)$ the probability measure

$$\mu(B_N) = \frac{1}{N}\sum_{i=1}^N \delta_{b_i^N}.$$

The sequence of fields B_N is said to converge in distribution to a probability measure μ if the sequence of measures $\mu(B_N)$ converges to μ as N goes to infinity. We suppose that there exists a probability measure μ on $\mathbb{R}$ such that :

• $(\mathbf{H_2})$: The sequence of fields $\overline{B}_N$ converges in distribution to μ as $N \to +\infty$.

Under assumption (H_1), this is equivalent to

• $(\mathbf{H_2'})$: The sequence of fields $\widetilde{B}_N$ converges in distribution to μ as $N \to +\infty$.

The last assumption states that the scale T_N used to define the averaged field $\overline{B}_N$ is small :

- $(\mathbf{H_3})$: $\lim_{N\to+\infty} \frac{r_N T_N}{N} = 0.$

We focus on configurations with at most r_N connected domains of denatured bonds, thus the mean length of a domain is of order N/r_N. Assumption (H_3) states that $T_N \ll N/r_N$. Averaging is done on a local scale. Note that we have three length scales : the global scale is of order N, the intermediate scale is of order N/r_N, and the local scale is of order T_N. In several applications, the length scale T_N will be a constant independent of N, that is why we use the term local scale.

Remark : In this paper, we focus on fields having values in $\{b_{AT}, b_{GC}\}$ because the external field stands for the DNA sequence. It is worth noting that we could work with general fields $B_N \in \mathbb{R}^N$. The results and proofs extend to this more general case under analogous but somewhat stronger hypotheses (essentially, we have to impose the convergence of the measures $\mu(B_N)$ to the measure μ and also the convergence of the first moment of $\mu(B_N)$ to the first moment of μ which is not automatic in the general case.)

Examples

In order to explain the hypotheses, we exhibit some sequences of external fields B_N satisfying $(H_1) - (H_3)$. The case of quasi-homogeneous fields and very-inhomogeneous fields were mentioned in [45].

Quasi-homogeneous fields. A sequence of fields B_N is said to be quasi-homogeneous with intensity b if it satisfies hypotheses $(H_1) - (H_3)$ with limit measure $\mu_1 = \delta_b$, for some $b \in \mathbb{R}$. Since the fields B_N only take values b_{AT} and b_{GC}, the parameter b must belong to $[b_{AT}, b_{GC}]$. This is a generalization of homogeneous fields : it is straightforward to check that if the fields are constant, for example with value b_{AT}, then hypotheses $(H_1) - (H_3)$ hold with limit measure $\mu = \delta_{b_{AT}}$ (take $T_N = 1$). Another example is that of periodic fields. Let W be a word in the letters $\{b_{AT}, b_{GC}\}$ of length T. We construct the fields B_N by repeating the word W (to have N letters, repeat it $[\frac{N}{T}]$ times and eventually add a few letters of the beginning of W). The mean intensity of the fields is approximatively $b = \frac{1}{T}\sum_{i\in W} b_i$. Take $T_N = T$. Because of periodicity, the averaged field $\overline{B}_N = (\bar{b}_i^N)_{1\le i\le N}$ is constant and equal to b, at least for $1 \le i \le N - T$. Take $\widetilde{B}_N$ be the constant field with value b and length N. It is easy to check that hypotheses $(H_1) - (H_3)$ are satisfied with the limit measure $\mu_1 = \delta_b$. We give a last example of quasi-homogeneous fields in the context of random external fields. Suppose that

- $(\mathbf{H_0'})$: There exists a sequence of integers T_N satisfying (H_3) such that for every $K > 0$,

the series $\sum_{N\geq 1} N\exp(-KT_N)$ converges.

This is stronger than assumption (H_0). Note that Hypothesis (H_0') is satisfied if $r_N = N^\delta$ for some $\delta \in (0,1)$, with $T_N = N^{(1-\delta)/2}$. Let $b \in [b_{AT}, b_{GC}]$. Suppose that the external field $B_N = (b_i^N)_{1\leq i\leq N}$ is random and corresponds to independent and identically distributed variables b_i^N, with values b_{AT} or b_{GC} and expectation b. Then almost every sequence of fields B_N is quasi-homogeneous with intensity b. We omit the proof since it is a consequence of Proposition 2.1 which will be proved in the sequel.

Very inhomogeneous fields. A sequence of fields B_N is said to be very inhomogeneous with intensity $b \in [b_{AT}, b_{GC}]$ if it satisfies hypotheses $(H_1) - (H_3)$ with limit measure

$$\mu_2 = \frac{b_{GC} - b}{b_{GC} - b_{AT}}\delta_{b_{AT}} + \frac{b - b_{AT}}{b_{GC} - b_{AT}}\delta_{b_{GC}}.$$

This measure is the one of greatest variance among the probability measures on $[b_{AT}, b_{GC}]$ with first moment b. This explains the terminology 'very inhomogeneous field'. Suppose the fields B_N belong to L_{Nr_N} and that $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N b_i^N = b$. There are large areas where the field B_N is constant (with values in $\{b_{AT}, b_{GC}\}$) and the mean intensity of the field is approximatively b. Then the sequence of fields B_N is very inhomogeneous with intensity b. To check it, take $T_N = 1$ and $\widetilde{B}_N = \overline{B}_N = B_N$. Hypothesis (H_1) is verified since B_N belongs to L_{Nr_N} and the convergence of the mean intensity to b ensures that hypothesis (H_2) holds with limit measure μ_2.

Random external fields. The random framework provides us an interesting family of examples.

Proposition 2.1. *Suppose that assumption (H_0') holds. Let $\beta : [0,1] \to [b_{AT}, b_{GC}]$ be a piecewise continuous function. For $N \geq 1$, define a random external field B_N such that the variables $b_i^N, 1 \leq i \leq N$ are independent and distributed according to*

$$\mathbb{P}(b_i^N = b_{GC}) = 1 - \mathbb{P}(b_i^N = b_{AT}) = \frac{\beta(i/N) - b_{AT}}{b_{GC} - b_{AT}}.$$

Then, $\mathbb{P}$-almost surely, the sequence of external fields (B_N) satisfies assumptions $(H_1) - (H_3)$ with limit distribution μ, which is the distribution of $\beta(U)$ for U uniformly distributed on $[0,1]$.

It is worth noting that a good choice of the function β allows us to obtain any limit measure μ on $[b_{AT}, b_{GC}]$.

Biological fields. What about biological DNA sequences ? In order to apply our results on denaturation to real biological DNA sequences, can we claim that our assumptions are satisfied ? The answer is far from being easy, and it is very striking to note that the question is of great biological interest. The first difficulty is that we generally consider only one DNA sequence and not a sequence of DNA sequences. Thus the asymptotic has to be interpreted as an approximation, which can be thought as effective since the length of DNA sequences is really high (for example several thousands pairs of bases for the DNA sequence of a virus, several billions pairs of bases for the human genome). What is the meaning of our assumptions ? The averaged field $\overline{B}_N$ is strongly connected to the GC-content along the sequence, which biologists plot with a moving window. Such plots are popular in genetics. Long domains with almost constant averaged field correspond to long parts of the sequence with homogeneous GC content. This strongly reminds us to the notion of isochore in genetics. An isochore is a long portion of a DNA sequence with homogeneous GC content. This notion was first mentioned by Bernardi [7]. Since then, several computational methods have been developed to exhibit the isochore structure in genome sequences (see for example the work of JL Olivier and al [47]). It appears that some sequences do exhibit such a structure and others do not. The notion of isochore is still a subject of research in genetics, and even of controversy, as show the most recent publications W.Li [40] or O.Clay and al [13].

2.2.2 The large deviations principle

At this point, we present the main mathematical result of the paper :

Theorem 2.1. *Assume that assumptions* $(H_0) - (H_3)$ *hold.*
Then the distribution of $(M_N, M_{NB_N}, R_N/N)$ *under the measure* π_{Nr_N} *satisfies a large deviations principle with speed* N *and good rate function* J_Δ *defined by*

$$J_\Delta(m, \tilde{m}, r) = \begin{cases} F(m, \tilde{m}) - \inf_\Delta F & \text{if } (m, \tilde{m}) \in \Delta \text{ and } r = 0 \\ +\infty & \text{otherwise} \end{cases}$$

where Δ *is the domain*

$$\Delta = \left\{ (m, \tilde{m}) \in \mathbb{R}^2 \mid 0 \leq m \leq 1 \ , \ \int_0^m F_\mu^{-1}(x)\, dx \leq \tilde{m} \leq \int_{1-m}^1 F_\mu^{-1}(x)\, dx \right\}. \tag{2.3}$$

The function F_μ^{-1} *denotes the pseudo-inverse of the repartition function of the probability measure* μ.

In this large deviations principle, the rate function J_Δ is defined in terms of the function F appearing in the definition of Benham's Hamiltonian, and in terms

of the domain Δ which catches all information about the field B_N. This domain will be of main importance in the sequel. In the quasi-homogeneous case with intensity b, the limit measure is $\mu_1 = \delta_b$ and the corresponding domain Δ_1 is the segment defined by

$$\Delta_1 = \{(m, \tilde{m}) \in \mathbb{R}^2 \mid 0 \leq m \leq 1 \ , \tilde{m} = bm \ \}.$$

In the very inhomogeneous case with intensity b, the limit measure is μ_2 and straightforward calculations show that the associated domain Δ_2 is the parallelogram defined by

$$\Delta_2 = \{(m, \tilde{m}) \in \mathbb{R}^2 \mid 0 \leq m \leq 1 \, , \max(b_{AT}m, b - b_{GC}(1-m)) \leq \tilde{m} \leq \min(b_{GC}m, b - b_{AT}(1-m))\}.$$

One diagonal of this parallelogram is the segment Δ_1. The two previous examples are extremal in the sense that they correspond to the most homogeneous and inhomogeneous fields respectively. The general shape of Δ satisfies the following :

Proposition 2.2. *For any distribution μ on $[b_{AT}, b_{GC}]$ with mean b, the domain Δ is convex, compact and symmetric around the point $(1/2, b)$. Furthermore, it contains the segment Δ_1 corresponding to the quasi-homogeneous case and is contained in the parallelogram Δ_2 corresponding to the very inhomogeneous case.*

2.2.3 The law of large numbers for denaturation

The asymptotic behavior of $(M_N, \widetilde{M}_{NB_N}, R_N/N)$ follows from the study of the good rate function J_Δ. We obtain the following :

Proposition 2.3. *Let $\kappa \neq \frac{4\pi^2 C}{K_0 A}$. When N goes to infinity, the distribution of $(M_N, \widetilde{M}_{NB_N}, R_N/N)$ under π_{Nr_N} converges in distribution to the Dirac measure at point $(M_\infty, \widetilde{M}_\infty, 0)$. The point $(M_\infty, \widetilde{M}_\infty)$ is the unique minimizer of the function F on Δ. It satisfies*

$$\widetilde{M}_\infty = \int_0^{M_\infty} F_\mu^{-1}(x)\, dx. \tag{2.4}$$

Equation (2.4) means that $\widetilde{M}_\infty$ lies on the lower boundary of Δ. Denaturation is localized so as to minimize $\widetilde{M}_\infty$. It occurs in regions where the averaged field $\overline{B}_N$ is low, i.e. regions with high AT concentration (since $b_{AT} < b_{GC}$).

In the last section, we use Proposition 2.3 to study denaturation as a function of superhelicity and give numerical applications.

2.3 Proof of the main results

2.3.1 Proof of Theorem 1

The strategy of the proof is the following : Varadhan's integral Lemma allows us to derive the LDP for the magnetization under the Gibbs measure π_{Nr_N} from a LDP for the magnetization under the uniform probability ρ_{N,r_N} formulated in Proposition 2.4. Thanks to the small perimeter assumption, the study of the magnetization under the uniform probability is reduced to combinatorial considerations formulated in Proposition 2.5. We have to study the existence of configurations with small perimeter and prescribed magnetization. The scale hypothesis (H_3) ensures that replacing the field B_N can be replaced by the field $\overline{B}_N$. Hypothesis (H_1) allows us to work with the field $\widetilde{B}_N$ instead of $\overline{B}_N$. It is helpful since $\widetilde{B}_N$ belongs to L_{Nr_N}. Hypothesis (H_2) ensures the convergence of the distribution of the fields and hence the convergence of several associated quantities.

Throughout this section, we suppose that $(B_N)_{N\geq 1}$ is a sequence of fields, (r_N) and (T_N) are sequences of integers, and that assumptions $(H_0) - (H_3)$ are satisfied.
We note $L = \max(|b_{AT}|, |b_{GC}|)$.

Reduction of the proof

Proposition 2.4. *The distribution of $(M_N, M_{NB_N}, R_N/N)$ under ρ_{N,r_N} satisfies a LDP with speed N and good rate function I_Δ defined by*

$$I_\Delta(m, \tilde{m}, r) = \begin{cases} 0 & \text{if } (m, \tilde{m}) \in \Delta \text{ and } r = 0 \\ +\infty & \text{otherwise} \end{cases} .$$

where Δ is the domain defined by (2.3).

Let us prove that Proposition 2.4 implies Theorem 2.1.
The Hamiltonian H_N defined by equation (2.1) is a function of M_N, M_{NB_N} and R_N only. Thus the distribution of $(M_N, M_{NB_N}, R_N/N)$ under the Gibbs measure π_{Nr_N} and under the uniform measure ρ_{N,r_N} are linked by the relation

$$\begin{aligned} &\int_{\Omega_N} \theta(M_N, M_{NB_N}, R_N/N) d\pi_{Nr_N} \\ =\ & 1/Z_{N,r_N} \int_{\Omega_N} \exp N[-aR_N/N - F(M_N, M_{NB_N})] \theta(M_N, M_{NB_N}, R_N/N) d\rho_{N,r_N} \end{aligned} ,$$

where $\theta : \mathbb{R}^3 \to \mathbb{R}$ is any bounded continuous function and Z_{N,r_N} is the partition function defined by

$$Z_{N,r_N} = \int_{\Omega_N} \exp N[-aR_N/N - F(M_N, M_{NB_N})] d\rho_{N,r_N}.$$

Since the function $(m, \tilde{m}, r) \mapsto -ar - F(m, \tilde{m})$ is bounded and continuous on $[0,1] \times [b_{AT}, b_{GC}] \times [0,1]$, we can apply Varadhan's integral Lemma to derive the LDP for the distribution of $(M_N, M_{NB_N}, R_N/N)$ under π_{Nr_N} from the LDP for the distribution of $(M_N, M_{NB_N}, R_N/N)$ under ρ_{N,r_N}. That is, Proposition 2.4 implies Theorem 2.1 thanks to Varadhan's integral Lemma. □

Lemma 2.1. *The following inequalities hold ρ_{N,r_N}-almost surely :*

$$|M_{NB_N} - M_{N\overline{B}_N}| \leq 2Lr_N T_N/N, \tag{2.5}$$

$$|M_{N\overline{B}_N} - M_{N\tilde{B}_N}| \leq \delta_N. \tag{2.6}$$

Thus the distribution of M_{NB_N}, $M_{N\overline{B}_N}$ and $M_{N\tilde{B}_N}$ under ρ_{N,r_N} are exponentially equivalent.

Proof : Let $\sigma \in \Omega_N$ be such that $R_N(\sigma) \leq r_N$. The subset $U = \{i | \sigma_i = +1\}$ of $\{1, .., N\}$ is the disjoint union of at most r_N 'connected components' $U_l = \{u_l, \cdots, v_l\}$. The difference is estimated by

$$|M_{NB_N}(\sigma) - M_{N\overline{B}_N}(\sigma)| \leq \frac{1}{N} \sum_{l=1}^{r_N} |\sum_{i \in U_l} b_i^N - \bar{b}_i^N|. \tag{2.7}$$

For the connected component U_l, we have

$$\sum_{i \in U_l} \bar{b}_i^N = \sum_{i=u_l}^{v_l} \frac{1}{T_N} \sum_{j=i}^{i+T_N-1} b_j = \sum_{i=u_l}^{v_l+T_N-1} \alpha_i b_i,$$

where $\alpha_i T_N$ is the cardinal of the set $\{i - T_N + 1, \cdots, i\} \cap \{u_l, \cdots, v_l\}$. If $u_l + T_N - 1 \leq i \leq v_l$, then $\alpha_i = 1$, otherwise $0 \leq \alpha_i \leq 1$. Therefore,

$$\left| \sum_{i \in U_l} b_i^N - \bar{b}_i^N \right| \leq \left| \sum_{i=u_l}^{u_l+T_N-1} (\alpha_i - 1) b_i + \sum_{i=v_l}^{v_l+T_N-1} \alpha_i b_i \right| \leq 2LT_N. \tag{2.8}$$

Equations (2.7) and (2.8) together yield equation (2.5).
Equation (2.6) is a straightforward consequence of assumption (H_1) :

$$|M_{N\overline{B}_N}(\sigma) - M_{N\tilde{B}_N}(\sigma)| \leq \frac{1}{N} \sum_{i=1}^{N} |\bar{b}_i^N - \tilde{b}_i^N| = \delta_N.$$

The notion of exponentially equivalent measures is defined in Dembo and Zeitouni ([19], p.130) : the distributions of M_{NB_N}, $M_{N\overline{B}_N}$ under ρ_{N,r_N} are said to be exponentially equivalent if for every $\alpha > 0$,

$$\limsup_{N \to \infty} \frac{1}{N} \log \rho_{N,r_N} \left(|M_{NB_N} - M_{N\overline{B}_N}| > \alpha \right) = -\infty. \tag{2.9}$$

Inequality (2.5) and hypothesis (H_3) imply that for large N,

$$\rho_{N,r_N}\left(|M_{NB_N} - M_{N\overline{B}_N}| > \alpha\right) = 0.$$

This implies equation (2.9). In the same way, inequality (2.6) and Hypothesis (H_1) imply that the distributions of $M_{N\overline{B}_N}$ and $M_{N\widetilde{B}_N}$ under ρ_{N,r_N} are exponentially equivalent. □

Two sequences of exponentially equivalent measures have the same large deviations properties, i.e. if a LDP hold for one sequence of measures, the same LDP will hold for any sequence of exponentially equivalent measures - see Theorem 4.2.13 in Dembo and Zeitouni [19]. Thus Lemma 1 allows us to work with $M_{N\widetilde{B}_N}$, instead of M_{NB_N}. It is equivalent to replace the field B_N by the field $\widetilde{B}_N$. It is helpful since $\widetilde{B}_N$ belongs to L_{Nr_N} and verifies Hypothesis (H_2'). It will be convenient to use simpler notations : in the sequel, we use the notation $\widetilde{M}_N$ instead of $M_{N\widetilde{B}_N}$.

Lemma 2.2. *Let $\omega_N \subset \Omega_N$ be a sequence of events such that for large N, $\rho_{N,r_N}(\omega_N) > 0$. Then*

$$\lim_{N\to+\infty} \frac{1}{N} \log[\rho_{N,r_N}(\omega_N)] = 0.$$

Proof : The probability measure ρ_{N,r_N} is the uniform probability on the set $\{\sigma \in \Omega_N | R_N(\sigma) \leq r_N\}$. As a configuration σ is uniquely determined by the value σ_1 and the position of the $2R_N(\sigma)$ sites i such that $\sigma_i \neq \sigma_{i+1}$, we have

$$\mathrm{card}\{\sigma \in \Omega_N | R_N(\sigma) \leq r_N\} = \sum_{r=0}^{r_N} 2\binom{N}{2r} \leq 2(r_N+1)\binom{N}{2r_N}.$$

The last inequality holds for large N, because if $4r_N \leq N$ and $0 \leq r \leq r_N$,

$$\binom{N}{2r} \leq \binom{N}{2r_N}.$$

If $\rho_{N,r_N}(\omega_N) > 0$, the following inequalities hold :

$$\left(2(r_N+1)\binom{N}{2r_N}\right)^{-1} \leq \rho_{N,r_N}(\omega_N) \leq 1.$$

We then use Stirling's formula to estimate the binomial coefficient and Hypothesis (H_0) to estimate the limit and the lemma is proved. □

Lemmas 2.1 and 2.2 allow us to reduce Proposition 2.4 to the following one :

Proposition 2.5.
• $(\mathbf{A_1})$: *For every closed set $A \subset \mathbb{R}^2$ disjoint from Δ, there exists $N_0 \in \mathbf{N}$ such that for every $N \geq N_0$, $\rho_{N,r_N}((M_N, \widetilde{M}_N) \in A) = 0$.*
• $(\mathbf{A_2})$: *For every open set $A \subset \mathbb{R}^2$ intersecting Δ, there exists $N_0 \in \mathbf{N}$ such that for every $N \geq N_0$, $\rho_{N,r_N}((M_N, \widetilde{M}_N) \in A) > 0$.*

Let us prove that Proposition 2.5 implies Proposition 2.4 and hence Theorem 2.1.
In view of Lemma 2.1, the LDP of Proposition 2.4 is equivalent to a LDP for the distribution of $(M_N, \widetilde{M}_N, R_N/N)$ under ρ_{N,r_N}, with good rate function I_Δ. We have to show that for every open set $O \subset \mathbb{R}^3$,

$$-\inf_{x\in O} I_\Delta(x) \leq \liminf_{N\to+\infty} \frac{1}{N} \log \rho_{N,r_N}\left((M_N, \widetilde{M}_N, R_N/N) \in O\right), \tag{2.10}$$

and that for every closed set $C \subset \mathbb{R}^3$,

$$\limsup_{N\to+\infty} \frac{1}{N} \log \rho_{N,r_N}\left((M_N, \widetilde{M}_N, R_N/N) \in C\right) \leq -\inf_{x\in C} I_\Delta(x). \tag{2.11}$$

As I_Δ equals 0 on the set $\Delta \times \{0\}$ and $+\infty$ outside of this set, inequalities (2.10) and (2.11) are trivial unless O is open and intersect $\Delta \times \{0\}$ or C is closed and disjoint from $\Delta \times \{0\}$. Let $O \in \mathbb{R}^3$ be an open set intersecting $\Delta \times \{0\}$. There exists an open set $A \subset \mathbb{R}^2$ intersecting Δ and $\varepsilon > 0$ such that $A\times]-\varepsilon, \varepsilon[\subset O$. Hypothesis (H_0) implies that for large N, $r_N/N < \varepsilon$ and thus

$$\rho_{N,r_N}((M_N, \widetilde{M}_N, R_N/N) \in O) \geq \rho_{N,r_N}((M_N, \widetilde{M}_N) \in A).$$

According to assertion (A_2), this probability is strictly positive for large N. Apply then Lemma 2.2 with $\omega_N = \{(M_N, \widetilde{M}_N) \in A\}$. This yields equation (2.10).
Let $C \in \mathbb{R}^3$ be a closed set disjoint from $\Delta \times \{0\}$. There exists a closed set $A \subset \mathbb{R}^2$ disjoint from Δ and $\varepsilon > 0$ such that $F \cap (\mathbb{R}^2 \times [-\varepsilon, \varepsilon]) \subset A \times [-\varepsilon, \varepsilon]$. Hypothesis (H_0) implies that for large N, $r_N/N < \varepsilon$ and thus

$$\rho_{N,r_N}((M_N, \widetilde{M}_N, R_N/N) \in C) \leq \rho_{N,r_N}((M_N, \widetilde{M}_N) \in C).$$

According to assertion (A_1), this probability is equal to zero for large N . Hence, the lim sup in equation (2.11) is equal to minus infinity, and this equation holds. □

Proof of Proposition 2.5

The following lemma investigates the behavior of $\widetilde{M}_N$ conditionally to M_N. Let $m_N \in [0, 1]$ be such that $Nm_N \in \mathbf{N}$. Define the set

$$\mathcal{V}_{N,m_N} = \{\widetilde{M}_N(\sigma) | R_N(\sigma) \leq r_N, M_N(\sigma) = m_N\}.$$

Lemma 2.3. *Then,*

$$\mathcal{V}_{N,m_N} = \{\widetilde{M}_N(\sigma) | M_N(\sigma) = m_N\} \tag{2.12}$$

Let $\tilde{m}$ be such that $\min \mathcal{V}_{N,m_N} \leq \tilde{m} \leq \max \mathcal{V}_{N,m_N}$. There exists $\tilde{m}' \in \mathcal{V}_{N,m_N}$ such that

$$|\tilde{m}' - \tilde{m}| \leq L/N. \tag{2.13}$$

Suppose that $m_N \to m$ as $N \to +\infty$. Then :

$$\lim_{N\to\infty} \min \mathcal{V}_{N,m_N} = \int_0^m F_\mu^{-1}(x)\, dx, \tag{2.14}$$

$$\lim_{N\to\infty} \max \mathcal{V}_{N,m_N} = \int_{1-m}^1 F_\mu^{-1}(x)\, dx. \tag{2.15}$$

Proof : Equation (2.12) states that we can forget the constraint $R_N \leq r_N$ in the definition of $\mathcal{V}_{N,m_N}$. This is true because the field $\widetilde{B}_N$ belongs to L_{Nr_N} - i.e. is constant on r_N subintervals. Divide $\{1, \cdots, N\}$ in r_N subintervals $U_1, \cdots, U_{r_N}$ where the field $\widetilde{B}_N$ is constant. Let $\sigma \in \Omega_N$ be a configuration such that $M_N(\sigma) = m_N$ and let J be the set of the Nm_N indexes where $\sigma = +1$. The magnetization $\widetilde{M}_N(\sigma)$ only depends on the cardinality of the sets $J \cap U_1, \cdots, J \cap U_{r_N}$. We can modify the configuration σ in the following way : choose two indexes i and $j \in U_l$ such that $\sigma_i = 1$ and $\sigma_j = 0$ and modify the configuration σ by setting $\sigma_i = 0$ and σ_j=1. This modification does not change the values $M_N(\sigma)$ and $\widetilde{M}_N(\sigma)$. It is possible to perform several modifications in such a way that all the elements of $J \cap U_i$ come on the left or on the right side of U_i. We obtain a configuration $\sigma' \in \Omega_N$ such that $M_N(\sigma') = M_N(\sigma) = m_N$, $\widetilde{M}_N(\sigma') = \widetilde{M}_N(\sigma)$ and $R_N(\sigma') \leq r_N$. This proves equation (2.12).
Let σ^- (resp. σ^+) be a configuration in $\{\sigma \in \Omega_N | M_N(\sigma) = m_N\}$ such that $\widetilde{M}_N$ is minimal (resp. maximal). We can find a path $\sigma_0 = \sigma^-, \cdots, \sigma_k = \sigma^+$ such that two successive configurations differ only by switching two spins, one from 0 to 1 and the other one from 1 to 0. Each configuration verifies $M_N(\sigma) = m_N$ and two successive values in the sequence $\widetilde{M}_N(\sigma_0), \cdots, \widetilde{M}_N(\sigma_l)$ differ by at most $2L/N$. This explains equation (2.13).
Let $\tilde{\mu}_N = \mu(\widetilde{B}_N) = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{b}_i^N}$ be the distribution of the external field $\widetilde{B}_N$. We denote by $F_{\tilde{\mu}_N}^{-1}$ the pseudo inverse of the repartition function of the measure $\tilde{\mu}_N$. It is a step function defined on $(0,1)$ which value on $]\frac{k-1}{N}, \frac{k}{N}]$ is the k-th smallest value among the $\tilde{b}_i^N$. Thus,

$$\min \mathcal{V}_{N,m_N} = \frac{1}{N}\left(F_{\tilde{\mu}_N}^{-1}(\frac{1}{N}) + \cdots + F_{\tilde{\mu}_N}^{-1}(\frac{Nm_N}{N})\right) = \int_0^{m_N} F_{\tilde{\mu}_N}^{-1}(x)\, dx. \tag{2.16}$$

Hypothesis (H_2') states that the sequence of measures $\tilde{\mu}_N$ converge to μ. This implies the convergence almost everywhere of the inverse repartition functions $F_{\tilde{\mu}_N}^{-1}$ to F_μ^{-1}. We also have the inequality $|F_{\tilde{\mu}_N}^{-1}(x)| \leq L$. Since $m_N \to m$ as $N \to +\infty$, Lebesgue's theorem implies that

$$\lim_{N\to+\infty} \int_0^{m_N} F_{\tilde{\mu}_N}^{-1}(x)\, dx = \int_0^m F_\mu^{-1}(x)\, dx. \tag{2.17}$$

Equations (2.16) and (2.17) together yield equation (2.14).
Equation (2.15) is proved in the same way. □

Proof of assertion (A_1) **:**
Suppose that the assertion does not hold : there exists a closed set A not intersecting Δ such that

$$\exists N_i \to +\infty,\ \exists \sigma_{N_i} \in \Omega_{N_i, r_{N_i}} \text{ such that } (M_{N_i}(\sigma_{N_i}), \tilde{M}_{N_i}(\sigma_{N_i})) \in A.$$

To simplify the notations, we omit the index i and write N instead of N_i. Since the magnetization $(M_N(\sigma), \widetilde{M}_N(\sigma))$ belongs to the compact set $[0,1] \times [b_{AT}, b_{GC}]$, we can assume that

$$(M_N(\sigma_N), \tilde{M}_N(\sigma_N)) \underset{N\to\infty}{\to} (m, \tilde{m}).$$

As A is a closed set, $(m, \tilde{m}) \in A$. Furthermore we have for each N,

$$\min \mathcal{V}_{N, M_N(\sigma_N)} \leq \widetilde{M}_N(\sigma_N) \leq \max \mathcal{V}_{N, M_N(\sigma_N)}.$$

Let N go to infinity, and use equations (2.14) and (2.15), this yields

$$\int_0^m F_\mu^{-1}(x)\, dx \leq \tilde{m} \leq \int_{1-m}^1 F_\mu^{-1}(x)\, dx.$$

It means that $(m, \tilde{m}) \in \Delta$. But $(m, \tilde{m})$ belongs to A and A doesn't intersect Δ. There is a contradiction and assertion (A_1) must hold. □

Proof of assertion (A_2) **:**
Let A be an open set intersecting Δ and let $(m, \tilde{m}) \in \Delta \cap A$. We exhibit a sequence $\sigma_N \in \Omega_N$ such that $R_N(\sigma_N) \leq r_N$ and

$$(M_N(\sigma_N), \widetilde{M}_N(\sigma_N)) \underset{N\to+\infty}{\longrightarrow} (m, \tilde{m}).$$

This will imply that for large N, $(M_N(\sigma_N), \widetilde{M}_N(\sigma_N))$ belongs to A and $\rho_{N, r_N}((M_N, \widetilde{M}_N) \in A) > 0$, proving assertion (A_2).
Let $m_N = [mN]/N$. This sequence verifies $m_N \underset{N\to+\infty}{\longrightarrow} m$.
If $\tilde{m}$ is such that $\min \mathcal{V}_{N,m_N} \leq \tilde{m} \leq \max \mathcal{V}_{N,m_N}$, then equation (2.13) implies that there exists $\sigma_N \in \Omega_N$ such that $R_N(\sigma_N) \leq r_N$, $M_N(\sigma_N) = m_N$ and $|\widetilde{M}_N(\sigma_N) - \tilde{m}| \leq L/N = \varepsilon_N^{(1)}$.
If $\tilde{m} < \min \mathcal{V}_{N,m_N}$, we choose σ_N such that $R_N(\sigma_N) \leq r_N$, $M_N(\sigma) = m_N$ and $\widetilde{M}_N(\sigma_N) = \min \mathcal{V}_{N,m_N}$. As $(m, \tilde{m}) \in \Delta$, we have the following inequalities

$$|\widetilde{M}_N(\sigma_N) - \tilde{m}| = \min \mathcal{V}_{N,m_N} - \tilde{m} \leq \min \mathcal{V}_{N,m_N} - \int_0^m F_\mu^{-1}(x)\, dx = \varepsilon_N^{(2)}.$$

In the same way, if $\tilde{m} > \max \mathcal{V}_{N,m_N}$, we prove a similar result with

$$\varepsilon_N^{(3)} = \int_{1-m}^1 F_\mu^{-1}(x)\, dx - \max \mathcal{V}_{N,m_N}.$$

The three cases together ensure that there exists $\sigma_N \in \Omega_N$ such that $R_N(\sigma_N) \leq r_N$, $M_N(\sigma_N) = m_N$ and $|\widetilde{M}_N(\sigma_N) - \tilde{m}| \leq \varepsilon_N$ with $\varepsilon_N = \max(\varepsilon_N^{(1)}, \varepsilon_N^{(2)}, \varepsilon_N^{(3)})$. Equations (2.14) and (2.15) imply that ε_N converges to 0 as N goes to infinity. As a consequence, $\widetilde{M}_N(\sigma_N) \underset{N\to+\infty}{\longrightarrow} \tilde{m}$ and assertion (A_2) is proved. □

2.3.2 Proof of Proposition 2.2

This proposition describes the general shape of Δ.

Proof : The function F_μ^{-1} is non decreasing and bounded. Thus the function $m \mapsto \int_0^m F_\mu^{-1}(x)\,dx$ is continuous, convex, bounded and $m \mapsto \int_{1-m}^1 F_\mu^{-1}(x)\,dx$ is continuous, concave and bounded. This proves that Δ is compact and convex. The relation

$$\int_0^{1-m} F_\mu^{-1}(x)\,dx = \int_0^1 F_\mu^{-1}(x)\,dx - \int_{1-m}^1 F_\mu^{-1}(x)\,dx = b - \int_{1-m}^1 F_\mu^{-1}(x)\,dx,$$

shows that Δ is symmetric around the point $(1/2, b)$. Since the points $(0,0)$ and $(1,b)$ belong to the convex domain Δ, the segment Δ_1 between these points is contained in Δ. Since F_μ^{-1} takes its values in $[b_{AT}, b_{GC}]$,

$$m b_{AT} \leq \int_0^m F_\mu^{-1}(x)\,dx \leq \int_{1-m}^1 F_\mu^{-1}(x)\,dx \leq m b_{GC}. \tag{2.18}$$

As $\int_0^1 F_\mu^{-1}(y)dy = b$,

$$b - b_{GC}(1-m) \leq \int_0^m F_\mu^{-1}(x)\,dx \leq \int_{1-m}^1 F_\mu^{-1}(x)\,dx \leq b - b_{AT}(1-m) \tag{2.19}$$

Equations (2.18) and (2.19) imply the inclusion $\Delta \subseteq \Delta_2$. □

2.3.3 Proof of Proposition 2.3

In this section, we show some applications of the large deviations principle stated in Theorem 1 : the magnetizations obey a law of large numbers, that is to say converge to a deterministic limit as the system size N goes to infinity. This result is derived from the study of the minima of the good rate function J_Δ. We assume that the hypotheses of Theorem 1 are satisfied with the limit measure μ. The Hamiltonian is defined in equation (2.1), it depends on physical constants a, b_{AT}, b_{GC}, C, K_0, A, and on the superhelicity κ.

Proof : The good rate function J_Δ has value $+\infty$ outside of $\Delta \times \{0\}$. On $\Delta \times \{0\}$, it is defined by

$$J_\Delta(m, \tilde{m}, 0) = F(m, \tilde{m}) - \inf_\Delta F = \frac{2\pi^2 C K_0}{4\pi^2 C + K_0 m}(\kappa + \frac{m}{A})^2 + \tilde{m} - \inf_\Delta F = G(m) + \tilde{m},$$

where G denote the function defined on $[0,1]$ by

$$G(m) = \frac{2\pi^2 C K_0}{4\pi^2 C + K_0 m}(\kappa + \frac{m}{A})^2 - \inf_{\Delta} F.$$

In order to minimize $J_\Delta(m, \tilde{m}, 0)$, we choose the lowest $\tilde{m}$, that is $\tilde{m} = \int_0^m F_\mu^{-1}(x)\, dx$, and minimize on $[0,1]$ the function ϕ defined by

$$\phi(m) = G(m) + \int_0^m F_\mu^{-1}(x)\, dx.$$

The function G has a second derivative given by

$$G''(m) = \frac{4\pi^2 C K_0 (K_0 \kappa A - 4\pi^2 C)^2}{(4\pi^2 C + K_0 m)^3 A^2}.$$

Hence for $\kappa \neq \frac{4\pi^2 C}{K_0 A}$, the function G is strictly convex. As F_μ^{-1} is non decreasing, its primitive is convex. Hence, for $\kappa \neq \frac{4\pi^2 C}{K_0 A}$, the function ϕ is strictly convex on $[0,1]$, and achieves its minimum at a unique point M_∞. Let $\widetilde{M}_\infty = \int_0^{M_\infty} F_\mu^{-1}(x)\, dx$. The rate function J_Δ achieves its minimum at a unique point which is $(M_\infty, \widetilde{M}_\infty, 0)$. The large deviations principle stated in Theorem 1 implies the convergence of the distribution of $(M_N, \widetilde{M}_N, R_N/N)$ under π_{Nr_N} to the Dirac measure at point $(M_\infty, \widetilde{M}_\infty, 0)$ with an exponential speed. □

Remark : If $\kappa = \frac{4\pi^2 C}{K_0 A}$, the good rate function J_Δ reduces on $\Delta \times \{0\}$ to the linear function

$$J_\Delta(m, \tilde{m}, 0) = \frac{\kappa}{2A}(\kappa + \frac{m}{A}) + \tilde{m} - \inf_{\Delta} F.$$

The uniqueness of a minimizer depends on the shape of Δ. In the case $0 < b_{AT} < b_{GC}$ (more important in applications), uniqueness always holds.

2.3.4 Proof of Proposition 2.1

In this section, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. For $N \geq 1$, the field B_N is a random variable from Ω to $\{b_{AT}, b_{GC}\}^N$. We denote by $B_N(\omega)$ a realization of the random variable B_N.
Let $\beta \to [b_{AT}, b_{GC}]$ be a piecewise continuous function. We suppose that for $N \geq 1$, the distribution of the random field B_N is such that the variables $b_i^N, 1 \leq i \leq N$ are independent and distributed according to

$$\mathbb{P}(b_i^N = b_{GC}) = 1 - \mathbb{P}(b_i^N = b_{AT}) = \frac{\beta(i/N) - b_{AT}}{b_{GC} - b_{AT}}.$$

The variable b_i^N is $\{b_{AT}, b_{GC}\}$-valued, with expectation $\beta(i/N)$.
Suppose furthermore that the sequence of integers r_N satisfies assumption

(H_0'). Proposition 2.1 states that $\mathbb{P}$-almost surely, the sequence of fields B_N satisfies assumptions $(H_1) - (H_3)$ with limit distribution μ, which is the distribution of $\beta(U)$ for U uniformly distributed on $[0,1]$.
The proof of Proposition 2.1 relies on a lemma of uniform exponential concentration for the empirical mean of Bernoulli variables :

Lemma 2.4. *Let $X_1, \cdots, X_n$ be independent variables, X_i being $\{0,1\}$-valued with expectation x_i. Let $\bar{X}$ be the mean of the X_i's, and $\bar{x}$ be the mean of the x_i's. For every $\varepsilon > 0$, there exists $K(\varepsilon) > 0$ depending on ε only, such that*

$$\mathbb{P}(|\bar{X} - \bar{x}| > \varepsilon) \leq 2e^{-K(\varepsilon)n}.$$

Proof : The proof of this lemma is a straightforward application of the standard concentration inequality for bounded martingales differences - see for example Dembo and Zeitouni [19] section 2.4.1. □

Proof of Proposition 2.1 : Choose T_N given by assumption (H_0'). For $1 \leq i \leq N$, the variable $\bar{b}_i^N$ is the empirical mean of T_N independent $\{b_{AT}, b_{GC}\}$-valued variables. The expectation of b_i^N is $\bar{\beta}_N(i/N)$, where $\bar{\beta}_N$ is the function defined by

$$\bar{\beta}_N(x) = \frac{1}{T_N} \sum_{k=i}^{i+T_N-1} \beta(x + k/N).$$

with the periodic boundary condition $\beta(x+1) = \beta(x)$.
We apply Lemma 2.4 : for every $\varepsilon > 0$,

$$\mathbb{P}(|\bar{b}_i^N - \bar{\beta}_N(i/N)| > \varepsilon) \leq 2e^{-T_N K'(\varepsilon)}$$

with $K'(\varepsilon) = K(\varepsilon/(b_{GC} - b_{AT}))$. We need this normalization because the b_i^N are not in $\{0,1\}$ but in $\{b_{AT}, b_{GC}\}$. As a consequence, we have the following estimation

$$\mathbb{P}(\max_{1 \leq i \leq N} |\bar{b}_i^N - \bar{\beta}_N(i/N)| > \varepsilon) \leq 2Ne^{-T_N K'(\varepsilon)}.$$

Since the series $\sum_{N \geq 1} 2Ne^{-T_N K'(\varepsilon)}$ is finite, Borel Cantelli's Lemma implies that $\mathbb{P}$-almost surely,

$$\max_{1 \leq i \leq N} |\bar{b}_i^N - \bar{\beta}_N(i/N)| \underset{N \to +\infty}{\longrightarrow} 0. \tag{2.20}$$

We now study the behavior of the deterministic functions $\bar{\beta}_N$ as N goes to infinity. If β is continuous on $[0,1]$ and $\beta(0) = \beta(1)$, then β is uniformly continuous on $[0,1]$ (with the boundary condition). As a consequence, the sequence of functions $\bar{\beta}_N$ uniformly converges on $[0,1]$ to β as N goes to infinity. If β is piecewise continuous on $[0,1]$, the uniform convergence holds on any compact where β is continuous, and the sequence of functions $\bar{\beta}_N$ converges to β in $L^1([0,1])$. This implies the convergence of the measures

$$\frac{1}{N} \sum_{i=1}^{N} \delta_{\bar{\beta}_N(i/N)} \underset{N \to +\infty}{\longrightarrow} \beta(U), \tag{2.21}$$

with U uniformly distributed on $[0,1]$.
Equations (2.20) and (2.21) implies that $\mathbb{P}$-almost surely

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\bar{b}_i^N} \underset{N\to+\infty}{\longrightarrow} \beta(U),$$

i.e. assumption (H_2) holds almost surely with limit measure μ equal to the distribution of the random variable $\beta(U)$.
As the piecewise continuous function β can be approximated by a sequence of piecewise constant functions, there exists $\widetilde{B}_N = (\tilde{b}_i^N)_{1\le i\le N} \in L_{Nr_N}$ such that

$$\frac{1}{N}\sum_{i=1}^{N} |\tilde{b}_i^N - \beta(i/N)| \underset{N\to+\infty}{\longrightarrow} 0. \tag{2.22}$$

Equations (2.22), (2.20) and the convergence in $L^1([0,1])$ of the sequence $\bar{\beta}_N$ to β implies that $\mathbb{P}$-almost surely

$$\frac{1}{N}\sum_{i=1}^{N} |\tilde{b}_i^N - \bar{b}_i^N| \underset{N\to+\infty}{\longrightarrow} 0.$$

It means that Hypothesis (H_1) holds $\mathbb{P}$-almost surely. □

2.4 Application : denaturation as a function of the superhelicity

We study denaturation as a function of superhelicity κ and present numerical computations. Note that similar computations can be done to study the denaturation as a function of the temperature. The values taken for the physical constants $b_{AT}, b_{GC} \cdots$ are those given by Clote and Backofen [14] recalled in the introduction. We plot the function $\kappa \mapsto M_\infty(\kappa)$ in the quasi-homogeneous case (Figure 1) and in the very inhomogeneous case (Figure 2), both with mean intensity $(b_{AT} + 2b_{GC})/3$. In both cases, we represent the denaturation M_∞ of a DNA polymer (on the y-axis) as a function of its superhelicity κ (on the x-axis). Both polymers consist in a large number of nucleotides, with a concentration in AT of 33.3%, and in GC of 66.7%. The difference between the two polymers lies in the repartition of the nucleotides. The first one has an homogeneous repartition of the nucleotides along the sequence whereas the second one contains large areas with only A and T, and large areas with G and C. Note that for the computation, we don't need the DNA sequence, but only the limit distribution μ. The computations are based on Proposition 2.3 : we compute the minimizer of the good rate function J_Δ on the domain $\Delta \times \{0\}$ in two different cases. In the homogeneous case, the limit distribution

is $\mu_1 = \delta_{0.333b_{AT}+0.667b_{GC}}$ and the domain Δ_1 is a segment. In the very inhomogeneous case, the limit distribution is $\mu_2 = 0.333\delta_{b_{AT}} + 0.667\delta_{b_{GC}}$ and the domain Δ_2 is a parallelogram.

We now comment upon these figures. In both cases, for $\kappa = 0$, the DNA polymer is not denatured - $M_\infty = 0$. This nondenatured state is stable : for $\kappa \approx 0$, we still have $M_\infty = 0$. On the other side, for large absolute value of the superhelicity $|\kappa|$, the DNA polymer is totally denatured - $M_\infty = 1$.

In the homogeneous case (Figure 1), the nondenatured state $M_\infty = 0$ is obtained for a superhelicity κ between the critical values $\kappa_1^- \approx -0.005$ and $\kappa_1^+ \approx 0.024$. If the superhelicity overcrosses the critical value κ_1^+, partial denaturation occurs - $M_\infty > 0$. The denaturation increases with the superhelicity, until it reaches the critical value $\kappa_2^+ \approx 0.175$ where the denaturation is total - $M_\infty = 1$. This totally denatured state is stable : for $\kappa \geq \kappa_2^+$, we still have $M_\infty = 1$. For negative values of the superhelicity, $\kappa_1^- \leq \kappa \leq 0$ correspond to the stable nondenatured state, $\kappa_2^- < \kappa < \kappa_1^-$ to partial denaturation and $\kappa \leq \kappa_2^-$ to the stable totally denatured state, with the critical value $\kappa_2^- \approx -0.156$. Note that $|\kappa_1^-| < \kappa_1^+$ and $|\kappa_2^-| < \kappa_2^+$: this reflects the fact that for a fixed amount of absolute superhelicity, the denaturation is larger for negative superhelicity than for positive superhelicity. In other terms, negative supercoiling enhances denaturation.

FIG. 2.1 – Denaturation M_∞ as a function of the superhelicity κ in the homogeneous case

In the very inhomogeneous case (Figure 2), the stable nondenatured state corresponds to a superhelicity between the critical values $\kappa_1^- \approx -0.001$ and

$\kappa_1^+ \approx 0.020$. The stable totally denatured state occurs for $\kappa \geq \kappa_2^+$ or $\kappa \leq \kappa_2^-$, with the critical values $\kappa_2^- \approx -0.173$ and $\kappa_2^+ \approx 0.192$. The intermediate values of the superhelicity $\kappa_2^- < \kappa < \kappa_1^-$ and $\kappa_1^+ < \kappa < \kappa_2^+$ yield partial denaturation. In this case, a new stable state appears corresponding to $M_\infty = 0.333$: this state corresponds to the complete denaturation of the AT domain and the non-denaturation of the GC domain. It occurs for $\kappa_3^+ \leq \kappa \leq \kappa_4^+$ and $\kappa_4^- \leq \kappa \leq \kappa_3^-$, with the critical values $\kappa_3^+ \approx 0.059$, $\kappa_3^- \approx -0.040$, $\kappa_4^+ \approx 0.081$ and $\kappa_4^- \approx -0.062$. For positive supercoiling, $0 \leq \kappa \leq \kappa_1^+$ correspond to the stable nondenatured state, for $\kappa_1^+ < \kappa < \kappa_3^+$ the AT domain is partially denaturated, for $\kappa_3^+ \leq \kappa \leq \kappa_4^+$ the stable state corresponding to the total denaturation of the AT domain is reached, for $\kappa_4^+ < \kappa < \kappa_2^+$ the GC domain is partially denatured and for $\kappa \geq \kappa_2^+$, the totally denatured state is reached. The same behavior holds for negative supercoiling. Once again, note that the critical values corresponding to negative supercoiling are smaller : for $i = 1, \cdots, 4$, $|\kappa_i^-| < \kappa_i^+$. This shows that negative supercoiling enhance denaturation.

FIG. 2.2 – Denaturation M_∞ as a function of the superhelicity κ in the very inhomogeneous case

Chapitre 3

A weighted random walk model. Application to a genetic algorithm.

Sommaire

Ce chapitre est tiré de l'article [22], soumis pour publication à *Journal of Applied Probability.*

Abstract

We consider a weighted random walk model defined as follows : the trajectory $(S_0, \cdots, S_n)$ of the symmetric simple random walk on the integers is given a probability proportional to $\prod_{k=1}^{n} f(S_k)$, the function f being the fitness function. We prove the convergence of the renormalized trajectory to a deterministic function with exponential speed. This function is a solution of a variational problem that we are able to solve explicitly. Our result is based on large deviations techniques and Varadhan's integral lemma.

We then study an application of this model to mutation-selection dynamics on the integers, where a random walk operates the mutation. These dynamics are the infinite population limit of mutation-selection genetic algorithms. We prove that the rate of escape of the population is of the order of the number of iteration steps and explicit the speed.

3.1 Introduction and motivations

3.1.1 The weighted random walk model

Let Ω_n be the set of nearest neighbour paths of length n :

$$\Omega_n = \left\{ S = (S_0, \cdots, S_n) \in \mathbb{Z}^{n+1} \mid \forall k = 1, \cdots, n,\ |S_k - S_{k-1}| = 1 \right\}.$$

Let $\mathbb{P}_{\pi,n}$ be the probability measure on Ω_n corresponding to the simple symmetric random walk with initial distribution π and let $\mathbb{E}_{\pi,n}$ be the associated expectation. Let $f : \mathbb{Z} \to [0, +\infty)$ be a fitness function. The weight of a path $S \in \Omega_n$ is defined by

$$\Pi_n^f(S) = \prod_{k=1}^{n} f(S_k).$$

The object of our study is the probability measure $\mathbb{P}_{\pi,n}^f$ on Ω_n defined by

$$\mathbb{P}_{\pi,n}^f(S) = \frac{1}{Z_{\pi,n}^f} \Pi_n^f(S) \mathbb{P}_{\pi,n}(S)$$

where $Z_{\pi,n}^f$ is the partition function

$$Z_{n,\pi}^f = \mathbb{E}_{\pi,n}[\Pi_n^f(S)]$$

that we suppose non zero. The associated expectation $\mathbb{E}^f_{\pi,n}$ verifies for any $\phi : \Omega_n \to \mathbf{R}$,

$$\mathbb{E}^f_{\pi,n}(\phi) = \frac{1}{Z^f_{\pi,n}} \mathbb{E}_{\pi,n}[\phi(S)\Pi^f_n(S)]. \tag{3.1}$$

In other words, the weighted random walk model gives to a path S a probability $\mathbb{P}^f_{\pi,n}(S)$ proportional to $\Pi^f_n(S)\ \mathbb{P}_{\pi,n}(S)$.

Remark 3.1. *The system $(\Omega_n, \mathbb{P}^f_{\pi,n})$ is not consistent in the sense that there is no measure on the set of infinite paths whose finite dimensional distributions are $\mathbb{P}^f_{\pi,n}$. Furthermore if the support of the fitness function f is the set of positive integers, then the probability $\mathbb{P}^f_{\pi,n}$ forces the paths to remain nonnegative.*

When the starting point is some deterministic integer x, we write $\mathbb{P}^f_{x,n}(S)$ instead of $\mathbb{P}^f_{\delta_x,n}(S)$. The case of power fitness function f will be important in the sequel and we write $\mathbb{P}^\beta_{\pi,n}$ when dealing with the fitness function

$$f(x) = x^\beta = \begin{cases} e^{\beta \ln(x)} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}.$$

One is interested in the asymptotic properties of $\mathbb{P}^f_{\pi,n}$ when the length n goes to infinity, that is its convergence (law of large numbers) and its fluctuations (central limit theorem). For the simple symmetric random walk starting at one, the mean $\mathbb{E}_{1,n}(S_n) = 1$, the variance $\mathbb{E}_{1,n}(S_n^2) = n$ and limit theorems are known. The random walk is said to be *diffusive*. In the case of power fitness functions, the asymptotic behaviour of the weighted random walk model is quite different since it appears to be *ballistic*, with speed depending on the parameter β. In the present work, we propose a functional convergence theorem available for all $\beta > 0$. The proof is based on a functional large deviations principle for the weighted random walk.

Let Ω be the space of càd-làg real valued functions defined on $[0,1]$ endowed with the topology of the uniform norm $\|.\|_\infty$, and let $\mathcal{AC}$ be the subspace of absolutely continuous functions. Let $\Psi_n : \Omega_n \to \Omega, S \mapsto \Psi_n(S)$ be the renormalisation map defined by

$$\Psi_n(S) : t \mapsto \begin{cases} \frac{1}{n} S_{[nt]+1} & \text{if } 0 \leq t < 1 \\ \frac{1}{n} S_n & \text{if } t = 1 \end{cases}.$$

Let $\widetilde{\mathbb{P}}_{\pi,n} = \mathbb{P}_{\pi,n} \circ \Psi_n^{-1}$ and $\widetilde{\mathbb{P}}^f_{\pi,n} = \mathbb{P}^f_{\pi,n} \circ \Psi_n^{-1}$ be the image probability measures on Ω. The large deviations properties of $\widetilde{\mathbb{P}}_{\pi,n}$ are well-known :

Theorem 3.1. (Mogulskii) *The sequence of probability measures $\widetilde{\mathbb{P}}_{\pi,n}$ satisfies a large deviations principle of speed n and good rate function $I : \Omega \to [0, +\infty]$ defined by*

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^\star(\dot{\phi}(t))\ dt & \textit{if } \phi \in \mathcal{AC} \textit{ and } \phi(0) = 0 \\ +\infty & \textit{otherwise} \end{cases},$$

where $\Lambda^\star : \mathbb{R} \to [0, +\infty]$ is the good rate function associated with the symmetric simple random walk defined by

$$\Lambda^\star(x) = \begin{cases} \frac{1-x}{2}\log(1-x) + \frac{1+x}{2}\log(1+x) & \text{if } x \in [-1,1] \\ +\infty & \text{otherwise} \end{cases} .$$

For a proof of this theorem, see [19].

Let us state the main result of this paper. Let $F : \Omega \to [-\infty, \infty)$ be the functional defined by

$$F(\phi) = \int_0^1 \log(\phi(t))\, dt.$$

Theorem 3.2. *Let $\beta > 0$ fixed.*

1. *The sequence of probability measures $\widetilde{\mathbb{P}}^\beta_{\pi,n}$ satisfies the following large deviations upper bound : for any closed set $A \subset \Omega$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \widetilde{\mathbb{P}}^\beta_{\pi,n}(A) \leq - \inf_{\phi\in A} \left\{ I - \beta F - \inf_{\phi\in\Omega}(I - \beta F) \right\} \tag{3.2}$$

   where I is defined in Theorem 3.1.

2. *The functional $I - \beta F$ has a unique minimizer on Ω denoted by ψ_β. The sequence of probability measures $\widetilde{\mathbb{P}}^\beta_{\pi,n}$ converges exponentially fast to δ_{ψ_β}. In other words, for any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that for n large enough,*

$$\widetilde{\mathbb{P}}^\beta_{\pi,n} \left(\|\phi - \psi_\beta\|_\infty > \varepsilon \right) \leq e^{-n\delta}.$$

3. *Let G_β be the function defined by*

$$G_\beta(y) = \int_0^y \frac{dx}{\sqrt{1 - x^{2\beta}}}.$$

   The function ψ_β is equal to

$$\psi_\beta(t) = \frac{1}{G_\beta(1)} G_\beta^{-1}\left(G_\beta(1)t\right).$$

For $\beta = 1$ and $\beta = 1/2$, the limit has a simple expression :

$$\psi_1(t) = \frac{2}{\pi}\sin\left(\frac{\pi}{2}t\right) \quad , \quad \psi_{1/2}(t) = t - \frac{1}{2}t^2 .$$

This theorem will be proved in the second section.

3.1.2 Infinite population genetic algorithm

Finite population genetic algorithms (GAs), introduced by Holland [33] are widely used in applications. They are relevant in many areas, for instance biology, computer science, optimisation ... The task of the GAs is to search a fitness landscape for maximal values. A population of individuals, considered as candidate solutions to the given problem is evolved under steps of mutation and steps of selection. The dynamics of the GA simulates, supposedly like in natural systems, the survival of the fittest among the individuals.

Despite their numerous heuristic successes, mathematical results describing the behaviour of GAs are rather sparse. Among the exceptions are R. Cerf [11, 12], Y. Rabinovich and A. Wigderson [49], C. Mazza and D. Piau [46], P. Del Moral and A. Guionnet [16, 17, 18], J. Bérard and A. Bienvenüe [5, 6].

Mathematically speaking, genetic algorithms are Markov chains on product space E^p, where E is the state space and $p \geq 1$ is the size of the running population. The dynamic is the combination of two basic operators : mutation and selection. Mutation is driven by an ergodic Markov transition kernel $Q(.,.)$ on E. Let $x = (x_i)_{1 \leq i \leq p} \in E^p$ be a population. The effect of mutation on x is modeled by the random choice of a new population with probability

$$Q(x_1, .) \otimes Q(x_2, .) \otimes \cdots \otimes Q(x_p, .).$$

Selection uses a function $f \geq 0$ on E, usually called the fitness. Given a population x, consider the probability measure π_x on $\{x_1, \cdots, x_p\}$ defined by

$$\pi_x = \frac{1}{< f(x) >} \sum_{i=1}^{p} f(x_i) \delta_{x_i}, \quad < f(x) > = \sum_{i=1}^{p} f(x_i).$$

The population after selection is randomly choosen in $\{x_1, \cdots, x_p\}^p$ with probability $\pi_x^{\otimes p}$.

P. Del Moral and A. Guionnet [16, 17] study the infinite population limit of mutation-selection algorithms, in connection with nonlinear filtering. In particular, these authors show that the empirical law on E of the population $x(n)$ at time n converges toward a deterministic measure μ_n when the population size p goes to infinity. They investigate also the large deviations properties of this convergence.

The sequence of limit measures μ_n forms a measure valued dynamical system :

$$\mu_{n+1} = T(\mu_n).$$

The operator $T = W \circ M$ is the combination of the mutation operator M and the selection operator W with fitness $f \geq 0$. The mutation operator is defined by $M : \mu \mapsto \mu Q$, where Q is the mutation kernel. For a given measure μ such

that $\mu(f) \in (0, +\infty)$, the selection operator W replaces μ by $W(\mu) = \hat{\mu}$ such that for any bounded function,

$$\hat{\mu}(g) = \frac{\mu(fg)}{\mu(f)}. \tag{3.3}$$

In this article, we focus on the following particular case. The state system is the set of integers $\mathbb{Z}$. The mutation is performed by a step of a simple symmetric random walk, i.e. $M(\mu) = \mu * \nu$, with $\nu = \frac{1}{2}(\delta_1 + \delta_{-1})$. The fitness function is a power function $f(x) = x^\beta$ vanishing on the set of negative integers. The finite population model is used in biology to describe the evolution of population of RNA viruses [53, 34]. Its long time behaviour is studied by J. Bérard and A. Bienvenüe [5, 6]. Mazza and Piau [46] investigate the infinite population model in the case $\beta = 1$ and prove a law of large numbers, a central limit theorem and a large deviations principle. More precisely, let $\mu_0 = \delta_1$ and let X_n be a random variable of law μ_n. Then X_n/n converges to $2/\pi$ a.s. and the reduced random variable

$$\frac{1}{\sqrt{n}}(X_n - \frac{2}{\pi}n)$$

converges in law to a centered Gaussian law of variance $4/\pi^2$. Their proof is based on a precise estimate of the Laplace transform of X_n based on series calculus. This method does not extend to the case of non linear fitness. As a by product of our results on the weighted random walks, we prove a law of large numbers available for all power fitness function.

Theorem 3.3. *In the mutation-selection dynamics on the integers with mutation based on the simple symmetric random walk and power fitness $f(x) = x^\beta$, the population at time n is located around the point $v_\beta n$. More precisely, $\frac{X_n}{n}$ converges almost surely to v_β. The speed is given by*

$$v_\beta = \left(\int_0^1 \frac{dx}{\sqrt{1 - x^{2\beta}}} \right)^{-1}.$$

This method is robust enough to deal with more general mutation operators. We are interested and working on other mutation operators : mutation can be induced by more general random walks, dynamic random walks... Dealing with other fitness function would also be very interesting. Central limit theorems are expected but not proved to our knowledge.

3.2 Proof of Theorem 3.2

The main result of Theorem 3.2 is the large deviations upper bound.
For every path $S \in \Omega_n$, the weight $\Pi_n^\beta(S)$ can be written

$$\Pi_n^\beta(S) = \prod_{i=1}^n S_i^\beta = \exp\left(\beta n \frac{1}{n} \sum_{i=1}^n \log\left(\frac{S_i}{n}\right) + \beta n \log(n) \right) = n^{\beta n} \exp\left(\beta n F(\Psi_n(S))\right)$$

and the probability $\widetilde{\mathbb{P}}^{\beta}_{\pi,n}$ verifies : for any Borel set $A \subset \Omega$,

$$\widetilde{\mathbb{P}}^{\beta}_{\pi,n}(A) = \frac{\int_A \exp(\beta n F(\phi))\ d\widetilde{\mathbb{P}}_{\pi,n}(\phi)}{\int_\Omega \exp(\beta n F(\phi))\ d\widetilde{\mathbb{P}}_{\pi,n}(\phi)}.$$

In this setting, Varadhan's integral lemma is the adequate tool to derive the large deviations properties of the measures $\widetilde{\mathbb{P}}^{\beta}_{\pi,n}$ from the large deviations properties of $\widetilde{\mathbb{P}}_{\pi,n}$ (see section 4.3 in [19]).
The difficulty comes from the irregularity of the functional F which is neither bounded nor continuous. Inequality (3.2) is a direct consequence of the following two inequalities :
for any closed set $A \subset \Omega$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \left[\int_A \exp(\beta n F(\phi))\ d\widetilde{\mathbb{P}}_{\pi,n}(\phi) \right] \leq - \inf_{\phi \in A} (I - \beta F)(\phi) \tag{3.4}$$

and

$$\liminf_{n\to\infty} \frac{1}{n} \log \left[\int_\Omega \exp(\beta n F(\phi))\ \ d\widetilde{\mathbb{P}}_{\pi,n}(\phi) \right] \geq - \inf_{\phi \in \Omega} (I - \beta F)(\phi). \tag{3.5}$$

For sake of clarity, we divide the proof into three parts.

3.2.1 Premilinary : some properties of the functional $I - \beta F$

The domain of a functional $J : \Omega \to (-\infty, +\infty]$ is the set where $J < +\infty$.

Lemma 3.1. *The functionals I, $-\beta F$ and $I - \beta F$ are lower semicontinuous and strictly convex on their domains.*

Proof : Since the functions $\Lambda^\star$ and $-\beta \log$ are strictly convex on their domains and lower semicontinuous so are the functionals I , $-\beta F$ and $I - \beta F$. □

Lemma 3.2. *In any closed set $A \subset \Omega$, there exists a point ϕ_0 such that*

$$I(\phi_0) - \beta F(\phi_0) = \inf_{\phi \in A} \{I(\phi) - \beta F(\phi)\}.$$

Proof : The domain of the function $\Lambda^\star$ is $[-1, 1]$. Hence, if $\phi \in \Omega$ verifies $I(\phi) < +\infty$, then almost surely $|\dot{\phi}(t)| \leq 1$, $\phi(t) \leq t$ and $F(\phi) \leq \int_0^1 \log(t) dt = -1$. Suppose that there exists $\phi \in A$ such that $I(\phi) - \beta F(\phi) \leq M < +\infty$. Then, $I(\phi) \leq M + \beta F(\phi) < +\infty$ and $F(\phi) \leq -1$. Hence the infimum of $I - \beta F$ over A is equal to the infimum over the set $A \cap \{I \leq M\}$ which is compact since A is closed and $\{I \leq M\}$ is compact. The lower semicontinuous functional $I - \beta F$ reaches its infimum over the compact set at some point $\phi_0 \in A$. □

A direct consequence of the two previous lemmas is the existence and unicity of the minimizer ψ_β of the functional $I - \beta F$ on Ω. The existence follows from Lemma 3.2 and the unicity is granted by the strict convexity stated in Lemma 3.1. The following lemma investigates some properties of the minimizer ψ_β.

Lemma 3.3. *The function ψ_β is absolutely continuous, nonnegative, nondecreasing and concave. It vanishes at 0, is strictly positive on $(0, 1]$ and satisfies $\dot{\psi}_\beta(0) > 0$.*

Proof : The domain of I is included in the set of absolutely continuous functions vanishing at O. The domain of $-\beta F$ is included in the set of nonnegative functions. The function ψ_β belongs to the domain of the functional $I - \beta F$ which is included in the set of absolutely continuous, nonnegative functions vanishing at 0.
Let $\tilde{\psi}_\beta$ be the nondecreasing function defined by

$$\tilde{\psi}_\beta(t) = \int_0^t |\dot{\psi}_\beta(u)| \, du.$$

Since the function $\Lambda^\star$ is even, $I(\psi_\beta) = I(\tilde{\psi}_\beta)$. Furthermore, $\tilde{\psi}_\beta \geq \psi_\beta$ so that $F(\tilde{\psi}_\beta) \geq F(\psi_\beta)$. Since ψ_β is the unique minimizer of $I - \beta F$, the functions $\tilde{\psi}_\beta$ and ψ_β are equal and ψ_β is nondecreasing.
Let $\hat{\psi}_\beta$ be the concave function defined by

$$\hat{\psi}_\beta(t) = \sup \left\{ \int_E \dot{\psi}_\beta(u) \, du \; ; \; E \subset [0,1] \; , \; |E| = t \right\}.$$

Then $I(\psi_\beta) = I(\hat{\psi}_\beta)$, and $\hat{\psi}_\beta \geq \psi_\beta$ so that $F(\hat{\psi}_\beta) \geq F(\psi_\beta)$. This implies the equality $\hat{\psi}_\beta = \psi_\beta$ and the concavity of ψ_β.
If there exists some point u such that $0 < u \leq 1$ and $\psi_\beta(u) = 0$, then $\psi_\beta = 0$ on $[0, u]$ and $F(\psi_\beta) = -\infty$. This contradict the fact that ψ_β minimizes $I - \beta F$. In the same way, if $\dot{\psi}_\beta(0) \leq 0$, the concavity implies that $\psi_\beta \leq 0$ and $F(\psi_\beta) = -\infty$. Contradiction. □

Lemma 3.4. *Let ψ be a function such that for any $t \in (0, 1]$, $\psi(t) < t$. The open set $B_\varepsilon = \{\phi \in \Omega \mid \phi > \psi \text{ on } [\epsilon, 1]\}$ verifies*

$$\inf_{\phi \in B_\varepsilon} \{I(\phi)\} \leq I(\psi).$$

Proof : The case $I(\psi) = +\infty$ is trivial. Suppose $I(\psi) < +\infty$. For any $\delta > 0$, let ϕ_δ be the function defined by

$$\phi_\delta(t) = \begin{cases} t & \text{if} \quad 0 \leq t \leq \delta \\ \psi(t) + \delta - \psi(\delta) & \text{if} \quad \delta \leq t \leq 1 \end{cases}.$$

The functions ϕ_δ belong to B_ε and a straightforward computation gives

$$I(\phi_\delta) = I(\psi) + \int_0^\delta \left[\Lambda^\star(1) - \Lambda^\star(\dot{\psi}(t)) \right] dt.$$

This quantity has limit $I(\psi)$ when $\delta \to 0$. Hence, $\inf_{\phi \in B_\varepsilon} \{I(\phi)\} \leq I(\psi)$. □

3.2.2 The large deviations upper bound

Proof of (3.4) : We apply Lemma 4.3.6 of [19]. Let $A \subset \Omega$ be a closed set. The functional $F_A : \Omega \to [-\infty + \infty)$ defined by

$$F_A(\phi) = \begin{cases} -\infty & \text{if} \quad \phi \notin A \\ F(\phi) & \text{if} \quad \phi \in A \end{cases}$$

is upper semicontinuous and the following tail condition holds :

$$\lim_{M\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \int_\Omega e^{nF_A(\phi)} \mathbf{1}_{\{F_A(\phi)\geq M\}} \, d\widetilde{\mathbb{P}}_{\pi,n}(\phi) = -\infty.$$

This is a consequence of the inequality $F_A(\phi) \leq \|\phi\|_\infty$ and of a similar tail condition for the norm $\|.\|_\infty$. The large deviations upper bound holds for $\widetilde{\mathbb{P}}_{\pi,n}$ with the good rate function $I : \Omega \to [0, +\infty]$. We can apply Lemma 4.3.6 of [19] which yields inequality (3.4). □

Proof of (3.5) : The function F is not lower semi-continuous, so we are not able to use Lemma 4.3.4 of [19] to establish the lower bound. Let $B_\varepsilon = \{\phi \in \Omega \mid \phi > \psi_\beta \text{ on } [\epsilon, 1]\}$. We have

$$\begin{aligned}
& \widetilde{\mathbb{E}}_{\pi,n}(\exp(\beta n F)) \\
= & \int_\Omega \exp(\beta n F(\phi)) \, d\widetilde{\mathbb{P}}_{\pi,n}(\phi) \\
\geq & \int_{B_\varepsilon} \exp(\beta n F(\phi)) \, d\widetilde{\mathbb{P}}_{\pi,n}(\phi) \\
= & \int_{B_\varepsilon} \exp\left(\beta n \int_0^1 \log\left(\phi(t)\right) dt\right) \, d\widetilde{\mathbb{P}}_{\pi,n}(\phi) \\
\geq & \exp\left(\beta n \int_\varepsilon^1 \log(\psi_\beta(t)) \, dt\right) \\
& \qquad \times \int_{B_\varepsilon} \exp\left(\beta n \int_0^\varepsilon \log(\phi(t)) \, dt\right) \, d\widetilde{\mathbb{P}}_{\pi,n}(\phi) \qquad (3.6)
\end{aligned}$$

In the last integral, let us condition on $\phi(\varepsilon)$ and apply the Markov property. This yields

$$\int_{B_\varepsilon} \exp\left(\beta n \int_0^\varepsilon \log(\phi(t)) \, dt\right) \, d\widetilde{\mathbb{P}}_{\pi,n} = \widetilde{\mathbb{E}}_{\pi,n}[\, Y_1 \, Y_2 \,] \qquad (3.7)$$

where

$$\begin{aligned}
Y_1 &= \widetilde{\mathbb{E}}_{\pi,n}[\, \mathbf{1}_{B_\varepsilon} \mid \phi(\varepsilon) \,], \\
Y_2 &= \widetilde{\mathbb{E}}_{\pi,n}[\, \exp(\beta n \int_0^\varepsilon \log(\phi(t)) \, dt) \mid \phi(\varepsilon) \,].
\end{aligned}$$

If $\phi(\varepsilon) \leq \psi_\beta(\varepsilon)$, then $Y_1 = 0$. Let us define

$$K_n^\varepsilon = \inf_{y \geq \psi_\beta(\varepsilon)} \widetilde{\mathbb{E}}_{\pi,n} \left[\exp\left(\beta n \int_0^\varepsilon \log(\phi(t))\ dt \right) \mid \phi(\varepsilon) = y \right].$$

Then, $Y_1\ Y_2\ \geq Y_1\ K_n^\varepsilon$ and

$$\widetilde{\mathbb{E}}_{\pi,n}[\ Y_1\ Y_2\ ] \geq K_n^\varepsilon\ \widetilde{\mathbb{E}}_{\pi,n}[\ Y_1\ ] = K_n^\varepsilon\ \widetilde{\mathbb{P}}_{\pi,n}(B_\varepsilon). \tag{3.8}$$

Equations (3.6), (3.7) and (3.8) together yield

$$\begin{aligned} & \liminf_{n\to\infty} \frac{1}{n} \log \widetilde{\mathbb{E}}_{\pi,n} \left(\exp(\beta n F)\right) \\ = & \liminf_{n\to\infty} \frac{1}{n} \log \left[\int_\Omega \exp(\beta n F(\phi))\ d\widetilde{\mathbb{P}}_{\pi,n}(\phi) \right] \\ \geq & \beta \int_\varepsilon^1 \log\ \psi_\beta(t)\ dt + \liminf_{n\to\infty} \frac{1}{n} \log\ \widetilde{\mathbb{P}}_{\pi,n}(B_\varepsilon) + \liminf_{n\to\infty} \frac{1}{n} \log K_n^\varepsilon. \end{aligned} \tag{3.9}$$

The set B_ε is open, so the large deviations lower bound (Theorem 3.1) and Lemma 3.4 imply that

$$\liminf_{n\to\infty} \frac{1}{n} \log \widetilde{\mathbb{P}}_{\pi,n}(B_\varepsilon) \geq - \inf_{\phi \in B_\varepsilon} I(\phi) \geq -I(\psi_\beta).$$

Letting $\varepsilon \to 0$, equation (3.9) yields

$$\liminf_{n\to\infty} \frac{1}{n} \log \widetilde{\mathbb{E}}_{\pi,n} \exp(\beta n F) \geq \beta F(\psi_\beta) - I(\psi_\beta) + \limsup_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log K_n^\varepsilon.$$

Let $y > \psi_\beta(\varepsilon) \neq 0$. The conditional expectation

$$\widetilde{\mathbb{E}}_{\pi,n} \left[\exp(\beta n \int_0^\varepsilon \log\ (\phi(t))\ dt) \mid \phi(\varepsilon) = y \right]$$

is underestimated by the contribution of the path ϕ_0 having increments $+1$ between times 0 and $([n\varepsilon] + ny)/2$ and increments -1 between times $([n\varepsilon] + ny)/2$ and $[n\varepsilon]$. For any $t \in [0, \varepsilon]$, $\phi_0(t) > \psi_\beta(\varepsilon)t$, hence

$$\widetilde{\mathbb{E}}_{\pi,n} \left[\exp\left(\beta n \int_0^\varepsilon \log\ (\phi(t))\ dt \right) \mid \phi(\varepsilon) = y \right] \geq \exp\left(\beta n \int_0^\varepsilon \log\ (\psi_\beta(\varepsilon)t)\ dt \right) 2^{-[n\varepsilon]}.$$

Thus,

$$\frac{1}{n} \log K_n^\varepsilon \geq \int_0^\varepsilon \log\ (\psi_\beta(\varepsilon)t)\ dt - \frac{[n\varepsilon]}{n} \log(2).$$

Finally, since $\psi_\beta(0) = 0$ and $\dot{\psi_\beta}(0) > 0$ (Lemma 3.3),

$$\limsup_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log(K_n^\varepsilon) = 0.$$

This proves inequality (3.5).

Inequalities (3.4) and (3.5) together imply that the sequence $(\widetilde{\mathbb{P}}^{\beta}_{\pi,n})_{n\in\mathbb{N}}$ satisfies the large deviations upper bound. The convergence of the sequence of measures $(\widetilde{\mathbb{P}}^{\beta}_{\pi,n})_{n\in\mathbb{N}}$ to δ_{ψ_β} with exponential speed is a consequence of this inequality. Let $A_\varepsilon = \{\ \phi \in \Omega \mid \|\phi - \psi_\beta\|_\infty \geq \varepsilon\}$. By Lemma 3.2, the infimum of $I - \beta F$ on the set A_ε is attained. Since ψ_β does not belong to A_ε, the infimum over A_ε is strictly greater than the infimum over Ω and

$$\limsup_{n\to\infty} \frac{1}{n} \log \widetilde{\mathbb{P}}^{\beta}_{\pi,n}(A_\varepsilon) < 0.$$

So there exists $\delta = \delta(\varepsilon) > 0$ such that for all n sufficiently large,

$$\widetilde{\mathbb{P}}^{\beta}_{\pi,n}(A_\varepsilon) \leq \exp(-n\delta).$$

□

3.2.3 Identification of the minimizer ψ_β

Let us define the function

$$\phi_\beta(t) = \frac{1}{G_\beta(1)} G_\beta^{-1}(G_\beta(1)t), \quad t \in [0,1]$$

where

$$G_\beta(y) = \int_0^y \frac{dx}{\sqrt{1-x^{2\beta}}}.$$

We are now going to prove that ϕ_β is equal to ψ_β, the unique minimizer of $I - \beta F$.

Lemma 3.5. *The function ϕ_β satisfies for any $t \in (0,1]$*

$$\frac{d}{dt}\left[\Lambda^{\star'}(\dot{\phi}_\beta(t))\right] + \frac{\beta}{\phi_\beta(t)} = 0, \tag{3.10}$$

and

$$\Lambda^{\star'}(\dot{\phi}_\beta(1)) = 0. \tag{3.11}$$

Proof : Let $p = G_\beta(1)$. The derivatives of ϕ_β are

$$\dot{\phi}_\beta(t) = \sqrt{1 - (p\phi_\beta(t))^{2\beta}},$$

$$\ddot{\phi}_\beta(t) = -\beta p^{2\beta} \phi_\beta(t)^{2\beta-1}.$$

After some computations, we can prove that equation (3.10) is equivalent to

$$\frac{1}{1-\dot{\phi}_\beta(t)^2}\ddot{\phi}_\beta(t)+\frac{\beta}{\phi_\beta(t)}=0.$$

Equation (3.11) holds since $\dot{\phi}_\beta(1)=0$ and $\Lambda^{\star\prime}(0)=0$. □

Remark 3.2. *Equation* (3.10) *is the Euler-Lagrange equation corresponding to the minimization of the functional* $I-\beta F$. *Note that any function of the form* $\phi(t)=\frac{1}{p}G_\beta^{-1}(pt)$ *is a solution of the Euler-Lagrange equation (3.10) and vanishes at* 0. *The value* $p=G_\beta(1)$ *is the one ensuring equation (3.11).*

We now prove that ϕ_β minimizes $I-\beta F$. Let $\phi\in\Omega$ and $0<t\leq 1$. By the convexity of $\Lambda^\star$,

$$\Lambda^\star(\dot{\phi}(t))\geq\Lambda^\star(\dot{\phi}_\beta(t))+\Lambda^{\star\prime}(\dot{\phi}_\beta(t))(\dot{\phi}(t)-\dot{\phi}_\beta(t)).$$

By the convexity of $-\beta\log$,

$$-\beta\log(\phi(t))\geq-\beta\log(\phi_\beta(t))-\frac{\beta}{\phi_\beta(t)}(\phi(t)-\phi_\beta(t)).$$

Adding the preceding two inequalities and integrating on $(0,1]$ yields

$$(I-\beta F)(\phi)\geq(I-\beta F)(\phi_\beta)+\int_0^1\left[\Lambda^{\star\prime}(\dot{\phi}_\beta(t))(\dot{\phi}(t)-\dot{\phi}_\beta(t))-\frac{\beta}{\phi_\beta(t)}(\phi(t)-\phi_\beta(t))\right]dt.$$

Using an integration by parts, we get that

$$\begin{aligned}&(I-\beta F)(\phi)-(I-\beta F)(\phi_\beta)\\ \geq\ &\left[\Lambda^{\star\prime}(\dot{\phi}_\beta(t))(\phi(t)-\phi_\beta(t))\right]_0^1-\int_0^1\left(\frac{d}{dt}[\Lambda^{\star\prime}(\dot{\phi}_\beta(t))]+\frac{\beta}{\phi_\beta(t)}\right)(\phi(t)-\phi_\beta(t))\,dt.\end{aligned}$$

By equation (3.10), the integral vanishes. By equation (3.11), the bracket vanishes also. Hence,

$$(I-\beta F)(\phi)\geq(I-\beta F)(\phi_\beta)$$

and ϕ_β minimizes $I-\beta F$. □

3.3 Application to a genetic algorithm

3.3.1 The relation between weighted random walk and mutation-selection dynamic

We consider mutation selection dynamics on a state space E. Recall that for a given fitness f, the selection operator W transforms a measure μ such that

$\mu(f) \in (0, +\infty)$ into the measure $W(\mu) = \hat{\mu}$ such that for any bounded function g,

$$\hat{\mu}(g) = \frac{\mu(fg)}{\mu(f)}.$$

The mutation is supposed to be induced by an irreducible markovian kernel $Q : M(\mu) = \mu Q$. We consider the measure valued dynamical system defined by

$$\left\{ \begin{array}{l} \mu_0 \\ \mu_{n+1} = W \circ M(\mu_n) \end{array} \right. .$$

A crucial remark is that one can write the law μ_n as follows :

Lemma 3.6. *Let $(Y_n)_{n\geq 0}$ be a Markov chain, with Q as transition kernel, and μ_0 as initial probability measure. Then, for any $x \in E$*

$$\mu_n(x) = \frac{\mathbb{E}_{\mu_0}\left(\mathbf{1}_{\{Y_n=x\}}\Pi_n^f\right)}{\mathbb{E}_{\mu_0}\left(\Pi_n^f\right)}, \quad \text{where} \quad \Pi_n^f = \prod_{k=1}^{n} f(Y_k).$$

Hence, for any bounded function g,

$$\mu_n(g) = \frac{\mathbb{E}_{\mu_0}\left(g(Y_n)\Pi_n^f\right)}{\mathbb{E}_{\mu_0}\left(\Pi_n^f\right)}.$$

Proof : One uses an induction over $n \geq 0$. For $n = 0$ the result holds since $\Pi_0^f = 1$. By the Markov property, for any $x \in E$,

$$\begin{aligned} & \mathbb{E}_{\mu_0}\left(\mathbf{1}_{\{Y_{n+1}=x\}}\ \Pi_{n+1}^f\right) \\ = & \sum_{y\in E} \mathbb{E}_{\mu_0}\left(\mathbf{1}_{\{Y_{n+1}=x\}}\ f(x)\ \Pi_n^f\ \mathbf{1}_{\{Y_n=y\}}\right) \\ = & f(x) \sum_{y\in E} \mathbb{E}_{\mu_0}\left(\Pi_n^f\ \mathbf{1}_{\{Y_n=y\}}\right) Q(y,x) \end{aligned}$$

By the definition of the selection and mutation operator, $\mu_{n+1}(x)$ is proportional to $f(x)\mu_n Q(x)$, with

$$\mu_n Q(x) = \sum_{y\in E} \mu_n(y) Q(y,x).$$

Hence the result at rank $n+1$ is a consequence of the result at rank n. □

We focus on the following case : the state space E is the integers set $\mathbb{Z}$, and the mutation kernel Q is the one operating in the simple symmetric random walk :

$$Q(x,y) = \left\{ \begin{array}{ll} 1/2 & \text{if } |x-y| = 1 \\ 0 & \text{otherwise} \end{array} \right. .$$

With the notation of lemma 3.6 and those stated in the introduction, the distribution of $(Y_0, Y_1, \cdots, Y_n) \in \Omega_n$ is $\mathbb{P}_{\mu_0,n}$. By the definition of the weighted random walk given by equation (3.1), and the formula of lemma 3.6, the following relation between the two models hold :

Lemma 3.7. *The measure at time n in the selection-mutation dynamic is the same as the distribution of the final value S_n in the weighted random walk model of length n. In other words, the measure μ_n is equal to the distribution of S_n under $\mathbb{P}^f_{\mu_0,n}$.*

Note that it is also the distribution of $n\phi(1)$ under $\tilde{\mathbb{P}}^f_{\mu_0,n}$.

3.3.2 Results on the mutation-selection dynamic

Let f be a power fitness function : $f(x) = x^\beta$ for some $\beta > 0$. We prove in that case the law of large numbers for the mutation-selection dynamic.

Proof : Thanks to lemma 3.7, Theorem 3.3 is a direct consequence of Theorem 3.2. By the contraction principle, a large deviations upper bound for the sequence μ_n can be deduced from the large deviations upper bound for $\tilde{\mathbb{P}}_{\mu_0,n}$. Or more direclty,

$$\mu_n(\ |t/n - \psi_\beta(1)| > \varepsilon\ ) = \mathbb{P}^\beta_{\mu_0,n}(\ |S_n/n - \psi_\beta(1)| > \varepsilon\ ) = \tilde{\mathbb{P}}^\beta_{\mu_0,n}(\ |\phi(1) - \psi_\beta(1)| > \varepsilon\ ).$$

Hence by Theorem 3.2, there exists $\delta = \delta(\varepsilon) > 0$ such that for n large enough,

$$\mu_n(\ |t/n - \psi_\beta(1)| > \varepsilon\ ) \leq \tilde{\mathbb{P}}^\beta_{\mu_0,n}(\ \|\phi - \psi_\beta\| > \varepsilon\ ) \leq e^{-n\delta}.$$

Let X_n be a random variable of law μ_n. The previous inequality implies that X_n/n converges almost surely to v_β. The speed $v_\beta = \psi_\beta(1)$ is given explicitly by

$$v_\beta = \left(\int_0^1 \frac{dx}{\sqrt{1 - x^{2\beta}}} \right)^{-1}.$$

□

Remark 3.3. *The linear growth might come as a surprise, since the mean position of the random walk, without selection, is of order $\sqrt{n}$. This puts the light on the strong effect of selection. In a loose sense, selection takes profit of the fluctuations that the mutation operator yields. This can be seen easily on Figure 1. In the case $\beta = 0$, there is no selection, and the speed is equal to zero. The vertical tangent at $\beta = 0$ indicates the strength of selection, even with very small parameter β. The limit $v_\infty = 1$ shows that a very strong selection forces almost every mutation step to be $+1$.*

FIG. 3.1 – Speed v_β as a function of the selection strength β.

Remark 3.4. *Our method is robust enough to deal with other mutation operator. Nevertheless, it is generally impossible to solve the optimization problem exactly, and we obtain only a law of large numbers, but the speed is not explicited. As an example, we have numerically computed the speed in the following case : linear fitness function* $f(x) = x$ *and mutation corresponding to a step of law* $\nu = \frac{1-\alpha}{2}\delta_{-1} + \frac{1+\alpha}{2}\delta_1$, *with* $\alpha \in (-1, 1)$. *The results appear in Figure 2. Note that in the case* $-1 < \alpha < 0$, *the random walk is transient with negative speed, but the strength of selection yields a positive speed* $v_\beta > 0$ *for the mutation-selection dynamic.*

Remark 3.5. *It would be of great interest to deal with other fitness function. For example, one can wonder what happens for a logarithm fitness function : one suspects the speed to be zero, and there might arise a new behaviour between the diffusive and the ballistic one.*

FIG. 3.2 – Speed v_α as a function of the mutation strength α.

Chapitre 4

Data Structures with Dynamical Random Transitions.

Sommaire

Ce chapitre est une version légèrement modifiée de l'article [23], écrit en collaboration avec N. Guillotin, B.Pinçon et R.Schott et accepté pour publication dans *Random Structure and Algorithms*

Abstract

We present a (**non-standard**) probabilistic analysis of dynamic data structures whose sizes are considered as dynamic random walks. The basic operations (insertion, deletion, positive and negative queries, batched insertion, lazy deletion, etc...) are time dependent random variables. This model is a (small) step towards the analysis of these structures when the distribution of the set of histories is **not uniform**. As an illustration, we focus on list structures (linear lists, priority queues and dictionaries) but the technique is applicable as well to more advanced data structures.

4.1 Introduction

The integrated time and space costs of sequences of operations on list structures have been estimated by Flajolet and al. [26] by combinatorial methods. Louchard [41] and Maier [44] presented two different probabilistic analyses of these dynamic data structures which led to the same conclusion : the integrated space and time costs of a sequence of n supported operations converge, as n goes to infinity, to Gaussian random variables. All the above mentioned results have been proved under a set of assumptions which constitute the so called Markovian model and assuming uniform distribution on the set of histories. Following the conclusions of Knuth [35] about deletions that preserve randomness, Louchard and al [43] have shown how to analyse dynamic data structures in the more realistic model proposed by Knuth. The maxima properties (value and position) of these data structures have been analysed by Louchard, Kenyon and Schott [42]. As discussed by Maier [44], the model of equiprobable histories is unrealistic and necessitates rejection. The main purpose of this paper is to derive the asymptotic properties of data structure sizes considered as dynamic random walks introduced and studied by the second author [30, 29, 31].
In Section 2 we give some preliminaries on dynamic data structures. In Section 3 we define the probabilistic model under consideration that is the dynamic random walk in any dimension then we recall some of their properties and prove a functional large deviation principle. In Sections 4, 5 and 6, the asymptotic behaviour of the dynamic data structures size is studied as well as the storage cost function one. Section 7 is devoted to the example of linear lists when the transformation is a rotation.

4.2 Preliminaries

A data type is a specification of the basic operations allowed together with its set of possible restrictions. The following data types are commonly used :
Stack : keys are accessed by position, operations are insertion I and deletion D but are restricted to operate on the key positioned first in the structure (the "top" of the stack).
Linear list : keys are accessed by position, operations are insertion I and deletion D without access restrictions (linear lists make it possible to maintain dynamically changing arrays).
Dictionary : keys belonging to a totally ordered set are accessed by value, all four operations I, D, Q^+, Q^- are allowed without any restriction. Q^+ represents a positive (successful) query (or search). Q^- stands for a negative (unsuccessful) query (or search).
Priority queue : keys belonging to a totally ordered set are accessed by value, the basic operations are I and D, deletion D is performed only on the key of minimal value (of "highest priority").
Symbol table : this type is a particular case of dictionary where deletion always operates on the key last inserted in the structure, only positive queries are performed.
A data organization is a machine implementation of a data type. It consists of a data structure which specifies the way objects are internally represented in the machine, together with a collection of algorithms implementing the operations of the data type.
Stacks are almost always implemented by arrays or linked lists .
Linear lists are often implemented by linked lists and arrays.
Dictionaries are usually implemented by sorted or unsorted lists ; binary search trees have a faster execution time and several balancing schemes have been proposed : AVL, 2-3 and red-black trees. Other alternatives are h-tables and digital trees.
Priority queues can be represented by any of the search trees used for dictionaries, more interesting are heaps, P-tournaments, leftist tournaments, binomial tournaments, binary tournaments and pagodas. One can also use sorted lists, and any of the balanced tree structures.
Symbol tables are special cases of dictionaries, all the known implementations of dictionaries are applicable here.

Definition 4.1. *A schema (or path) is a word*

$$\omega = O_1 O_2 \dots O_n \in \{I, D, Q^+, Q^-\}^*$$

such that for all k, $1 \leq k \leq n$ *:*

$$|O_1 O_2 \dots O_k|_I \geq |O_1 O_2 \dots O_k|_D.$$

$\{I, D, Q^+, Q^-\}^*$ is the free monoid generated by the alphabet

$$\{I, D, Q^+, Q^-\}$$

$|\omega|_O$ is the number of O in the word ω.

A schema is to be interpreted as a sequence of n requests (the keys operated on not being represented).

Example
The figure below shows a schema.

FIG. 4.1 – A schema

Definition 4.2. *A structure history is a sequence of the form :*

$$h = O_1(r_1)O_2(r_2)\dots O_n(r_n)$$

where $\omega = O_1O_2\dots O_n$ is a schema, and the r_k are integers satisfying :
$1 \leq r_k \leq poss(O_k, \alpha_{k-1}(\omega))$ and $\alpha_k(\omega) = |O_1O_2\dots O_k|_I - |O_1O_2\dots O_k|_D$ is the size (level) of the structure at step k, poss is a possibility function defined on each request, r_k is the rank (or position) of the key operated upon at step k.

We will only consider schemas and histories with initial and final level 0.
Possibility functions.
Two different models have been considered for defining possibility functions : the markovian model [26, 41] in which possibility functions are linear functions of the size α of the data structure when an allowed operation is performed. Knuth's model is related to his observation [35] that deletions may not preserve randomness and is more realistic than the markovian model. The following simple example may be helpful to understand Knuth's fundamental remark. Consider again the sequence of operations $IIDI$ performed, for example, on a linear list which is initially empty. Let $x < y < z$ be the three keys inserted during the sequence III. x, y and z are deleted with equal probability. Let w be the key inserted by the fourth I. Then all four cases $w < x < y < z$, $x < w < y < z$, $x < y < w < z$, $x < y < z < w$ do occur with equal probability, whatever the key deleted. More generally, let us consider a sequence of operations $O_1O_2\dots O_k$ on a dictionary data type, the initial data structure being empty (any data type listed above may be considered). Assume O_k is

the ith I or Q^- of the sequence. Let $x_1 < x_2 < \ldots < x_{i-1}$ be the keys inserted and negatively searched during the sequence $O_1 O_2 \ldots O_{k-1}$, and let w be the ith inserted or negatively searched key. Then all the cases $w < x_1 < x_2 < \ldots < x_{i-1}$, $x_1 < w < x_2 < \ldots < x_{i-1}$, $\ldots$, $x_1 < x_2 < \ldots < x_{i-1} < w$ are equally likely, whatever the deleted keys. Put into combinatorial words : after k operations, whose i are I and Q^-'s (thus $k-i$ are D and Q^+'s), the size of the data structure is $\alpha \leq 2i - k$. The keys of the data structure can be considered as a subset of α distinct objects of a set of size i any of the C_i^α possible subsets being equally likely. We say that the number of possibilities of the ith I or Q^- (in a sequence of operations) is equal to i in Knuth's model whatever the size of the data structure when this insertion (or negative query) occurs. We summarize in the two tables below the differences between the markovian and Knuth's models. We consider only a few data structures.

Data type	poss(I,α)	poss(D,α)	poss(Q^+,α)	poss(Q^-,α)
Dictionary	$\alpha + 1$	α	α	$\alpha + 1$
Priority queue	$\alpha + 1$	1	0	0
Linear list	$\alpha + 1$	α	0	0

TAB. 4.1 – Possibility functions in the markovian model.

Data type	poss(i^{th} I)	poss(D,α)	poss(Q^+,α)	poss($i^{th}Q^-$)
Dictionary	i	α	α	i
Priority queue	i	1	0	0
Linear list	i	α	0	0

TAB. 4.2 – Possibility functions in Knuth's model

If instead of deleting items from the data structure as described above we wait until there is a new insertion, we get a new operation called lazy deletion (see [42] and the references therein). Batched insertion waits until there is a new deletion. The probabilistic model described below applies also for these operations but we restrict our analysis to dynamic data structures subject to the classical operations : I, D, Q^+, Q^-.

4.3 The probabilistic model

4.3.1 Definition

Let $S = (E, \mathcal{A}, \mu, T)$ be a dynamical system where $(E, \mathcal{A}, \mu)$ is a probability space and T is a measure-preserving transformation defined on E. Let $d \geq 1$ and $(e_j)_{1 \leq j \leq d}$ be the unit coordinate vectors of $\mathbb{Z}^d$. Let $f_1, \ldots, f_d$ be measurable

functions defined on E with values in $[0, \frac{1}{d}]$. For each $x \in E$, we denote by $\mathbb{P}_x$ the distribution of the time-inhomogeneous random walk :

$$S_0 = 0, \quad S_n = \sum_{i=1}^{n} X_i \quad \text{for} \quad n \geq 1$$

with step distribution

$$\mathbb{P}_x(X_i = z) = \begin{cases} f_j(T^i x) & \text{if} \quad z = e_j \\ \frac{1}{d} - f_j(T^i x) & \text{if} \quad z = -e_j \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

It is worth remarking that if the functions f_j are not all constant, $(S_n)_{n\in\mathbb{N}}$ is a non-homogeneous Markov chain. This Markov chain can be classified in the large class of random walks evolving in a random environment. In most of the papers (see for instance [28], [15],...), the environment field takes place in space but it can also take place in space and time (see [10]). Following the formalism used in the study of these random walks, when x is fixed, the measure $\mathbb{P}_x$ is called *quenched* and the measure averaged on values of x defined as $\mathbb{P}(.) = \int_E \mathbb{P}_x(.)\ d\mu(x)$ is called *annealed*. The dynamic random walks were introduced and studied by the second author in [30, 29, 31]. Let us recall the results already obtained in the quenched setting. Denote by $A = (a_{ij})_{1\leq i,j\leq d}$ the matrix with coefficients

$$\begin{aligned} a_{jj} &= \frac{1}{d^2}\int_E (d + 1 - 4d^2 f_j^2) d\mu \\ a_{ij} &= a_{ji} = \frac{1}{d^2}\int_E (1 - 4d^2 f_i f_j) d\mu. \end{aligned}$$

Let $\mathcal{C}_1(S)$ be the class of functions $f \in L^1(\mu)$ satisfying the condition : for μ-almost every point $x \in E$,

$$\left| \sum_{k=1}^{n} \Big(f(T^k x) - \int_E f d\mu \Big) \right| = o\Big(\frac{\sqrt{n}}{\log n} \Big).$$

Choose $f_j \in \mathcal{C}_1(S), j = 1, \ldots, d$ such that for every $j, l \in \{1, \ldots, d\}, f_j f_l \in \mathcal{C}_1(S)$ and $\int_E f_j\ d\mu = \frac{1}{2d}$. Then, for μ-almost every point $x \in E$, S_n satisfies a local limit theorem, namely

$$\mathbb{P}_x(S_{2n} = 0) \sim \frac{2}{\sqrt{\det A}(4\pi n)^{\frac{d}{2}}} \quad \text{as} \quad n \to \infty$$

(see [31]).
Let $\mathcal{C}_2(S)$ denote the class of functions $f \in L^1(\mu)$ satisfying the following condition :

$$\sup_{x\in E} \left| \sum_{i=1}^{n} \Big(f(T^i x) - \int_E f d\mu \Big) \right| = o\Big(\sqrt{n} \Big).$$

Assume that for every $j,l \in \{1,\dots,d\}$, $f_j \in \mathcal{C}_2(S)$, $f_j f_l \in \mathcal{C}_2(S)$ and $\int_E f_j d\mu = \frac{1}{2d}$, then, for every $x \in E$, the sequence of processes $(\frac{1}{\sqrt{n}} S_{[nt]})_{t\geq 0}$ weakly converges in the Skorohod space $\mathcal{D} = \mathcal{D}([0,\infty[)$ (see [9]) to the d-dimensional Brownian motion $B_t = (B_t^{(1)},\dots,B_t^{(d)})$ with zero mean and covariance matrix At (see [32]). This result was used to study some problems related to resource sharing, namely distributed algorithms as the well-known colliding stacks problem or the Banker's algorithm in the case when the requests are time dependent (see [32] for further details). Let us also remark that the dynamic $\mathbb{Z}^d$-random walks quite differ from the standard $\mathbb{Z}^d$-random walks on nearest neighbors since the matrix A is not necessarily diagonal and some dimensional correlations appear in the limit process $(B_t)_{t\geq 0}$. A strong law of large numbers for the dynamic random walks can be obtained for μ-almost every point $x \in E$ from Kolmogorov's theorem assuming that the functions $f_1,\dots,f_d$ are measurable. The limit vector is then given by $(2\mu(f_j|\mathcal{I}) - 1)_{1\leq j\leq d}$ where $\mathcal{I}$ is the invariant σ-field associated to the transformation T. So, $(S_n/n)_{n\geq 1}$ is a good candidate for the LDP. In the next section, we derive a functional LDP for the dynamic $\mathbb{Z}^d$-random walks.

4.3.2 Large Deviations Principles

Let Γ be a Polish space endowed with the Borel σ-field $\mathcal{B}(\Gamma)$. A good *rate* function is a lower semi-continuous function $\Lambda^* : \Gamma \to [0,\infty]$ with compact level sets $\{x; \Lambda^*(x) \leq \alpha\}, \alpha \in [0,\infty[$. Let $v = (v_n)_n \uparrow \infty$ be an increasing sequence of positive reals. A sequence of random variables $(Y_n)_n$ with values in Γ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to satisfy *a Large Deviation Principle* (LDP) with speed $v = (v_n)_n$ and the good rate function Λ^* if for every Borel set $B \in \mathcal{B}(\Gamma)$,

$$\begin{aligned} -\inf_{x\in B^o} \Lambda^*(x) &\leq \liminf_n \frac{1}{v_n} \log \mathbb{P}(Y_n \in B) \\ &\leq \limsup_n \frac{1}{v_n} \log \mathbb{P}(Y_n \in B) \leq -\inf_{x\in \bar{B}} \Lambda^*(x). \end{aligned}$$

In the following, $f_1,\dots,f_d$ are functions defined on E with values in $[0,\frac{1}{d}]$. We define the family $(l_\lambda)_{\lambda\in\mathbb{R}^d}$ of functions defined on E with values in $\mathbb{R}$ by :

$$l_\lambda := \log\Big(\sum_{j=1}^d (e^{\lambda_j} f_j + (\frac{1}{d} - f_j) e^{-\lambda_j})\Big).$$

For every λ, the function l_λ is bounded by $\log(\frac{2}{d}(\sum_{j=1}^d \cosh\lambda_j))$. It is measurable (resp. continuous) as soon as the functions (f_j) are measurable (resp. continuous).

Theorem 4.1. *For μ-a.e. $x \in E$, the distributions of S_n/n under $\mathbb{P}_x$ satisfy the LDP with speed n and the good rate function*

$$\Lambda^*(y) = \sup_{\lambda \in \mathbb{R}^d} \{< \lambda, y > -\Lambda(\lambda)\}$$

where

$$\Lambda(\lambda) = \mu(l_\lambda |\ \mathcal{I}),$$

$\mathcal{I}$ being the σ-field generated by the fixed points of the transformation T.

Let us assume E to be a compact metric space, $\mathcal{A}$ the associated Borel σ-field and T a continuous transformation of E. If there exists an unique invariant measure μ i.e. $(E, \mathcal{A}, \mu, T)$ is uniquely ergodic and if $f_1, \ldots, f_d$ are continuous, then the assertion of Theorem 4.1 holds for every $x \in E$. In that case,

$$\Lambda(\lambda) = \mu(l_\lambda).$$

The rate function is deterministic. Under these stronger hypotheses on the dynamical system, we can extend Theorem 4.1 as follows.
Let us define for every v$n \geq 1$,

$$S_n^*(t) := \frac{S_{[nt]}}{n}, \quad t \in [0,1].$$

The linear interpolation of $S_n^*(t)$, $t \in [0,1]$, is then defined by

$$\bar{S}_n(t) := S_n^*(t) + \Big(t - \frac{[nt]}{n}\Big) X_{[nt]+1}.$$

Let $\mathbb{P}_{x,n}$ and $\bar{\mathbb{P}}_{x,n}$ be the distribution of $S_n^*(.)$ and $\bar{S}_n(.)$ in $L_\infty([0,1])$ equipped with the supremum norm. Throughout, $\mathcal{C}([0,1])$ denotes the space of continuous functions on $[0,1]$ and $\mathcal{AC}([0,1])$ denotes the space of absolutely continuous functions on $[0,1]$.

Theorem 4.2. *Let $(E, \mathcal{A}, \mu, T)$ be an uniquely ergodic dynamical system. If $f_1, \ldots, f_d$ are continuous, then for every $x \in E$, the distributions $(\bar{\mathbb{P}}_{x,n})_{n \geq 1}$ satisfy in $\mathcal{C}([0,1])$ equipped with the supremum norm, the LDP with the good rate function*

$$I(\phi(.)) = \begin{cases} \int_0^1 \Lambda^*(\dot{\phi}(t))\ dt, & \text{if } \ \phi(.) \in \mathcal{AC}, \phi(0) = 0 \\ +\infty, & \text{otherwise.} \end{cases}$$

Remark : Let us mention that an annealed large deviations statement can easily be proved using results of [20] (E is assumed to be compact). Details are omitted since the dynamic operations on the data structures will be modeled from the quenched probability measure.

4.3.3 Proof of Theorem 4.1

By independence of (X_i), we have, for every $\lambda \in \mathbb{R}^d$ and every $x \in E$,

$$\frac{1}{n} \log \mathbb{E}_x(e^{<\lambda, S_n>}) = \frac{1}{n} \sum_{i=1}^{n} l_\lambda(T^i x).$$

Therefore the Birkoff's theorem implies that for μ-a.e. $x \in E$ and every $\lambda \in \mathbb{R}^d$,

$$\frac{1}{n} \log \mathbb{E}_x(e^{<\lambda, S_n>}) \to \mu(l_\lambda |\, \mathcal{I}) = \Lambda(\lambda).$$

The Λ being finite and differentiable on $\mathbb{R}^d$, by Gärtner-Ellis Theorem (see [19]), the theorem follows.

4.3.4 Proof of Theorem 4.2

The proof follows the same lines of argument that in [19] section 5.1. It is a Mogulskii-like theorem. The only difference lays in the proof of the LDP for the finite dimensional marginals, which we claim now.

Proposition 4.1. *Let $\mathcal{P}$ be the set of all ordered finite subsets of the interval $]0,1]$ that is the set of k-tuples $t^k = \{t_0 = 0 < t_1 < t_2 < \ldots < t_k \leq 1\}$ with $k \geq 1$. Let $f : [0,1] \to \mathbb{R}^d$; for every $k \geq 1$, for every k-tuple t^k, we define*

$$p_{t^k}(f) = (f(t_1), \ldots, f(t_k)) \in (\mathbb{R}^d)^k.$$

Then, the laws $\bar{\mathbb{P}}_{x,n} \circ p_{t^k}^{-1}$ satisfy in $(\mathbb{R}^d)^k$ the LDP with the good rate function

$$\Lambda_k^*(y) = \sum_{l=1}^{k} (t_l - t_{l-1}) \Lambda^* \left(\frac{y_l - y_{l-1}}{t_l - t_{l-1}} \right).$$

Proof :
The difference with the classical proof (see [19], Lemma 5.1.8., p. 178) is that the increments of the dynamic random walk are not stationary. But we can remark that for every $l \geq 1$, an increment of the dynamic random walk $S_{[nt_l]} - S_{[nt_{l-1}]}$ is a new dynamic random walk associated to the dynamical system $(E, \mathcal{A}, \mu, T)$, to the functions $f_1, \ldots, f_d$ and to the point $T^{[nt_{l-1}]}x$. Then the hypothesis of unique ergodicity on the dynamical system and Theorem 4.1 permits us to deduce the LDP for the laws $\mathbb{P}_{x,n} \circ p_{t^k}^{-1}$. □

4.3.5 A riemannian dynamic random walk

Let $f_1, \ldots, f_d$ be one-periodic functions defined on $\mathbb{R}$ with values in $[0, \frac{1}{d}]$ and $(X_{i,n})_{1 \leq i \leq n}$ be a sequence of independent random vectors with values in $\mathbb{Z}^d$ with distribution

$$\mathbb{P}_x(X_{i,n} = z) = \begin{cases} f_j(x + \frac{i}{n}) & \text{if } \ z = e_j \\ \frac{1}{d} - f_j(x + \frac{i}{n}) & \text{if } \ z = -e_j \\ 0 & \text{otherwise} \end{cases}$$

We write

$$S_0 = 0, \quad S_n = \sum_{i=1}^{n} X_{i,n} \text{ for } n \geq 1$$

this n-dynamic random walk. This random walk is more difficult to study due to the presence of n in the transition probabilities which creates much more temporal inhomogeneity. The proof of the following proposition is straightforward.

Proposition 4.2. *Let f be a one-periodic function which can be expanded into a Fourier series $f(x) = \sum_{h \in \mathbb{Z}} c_h e^{2\pi i h x}$.*
When there exists $\beta > 1$ such that $|c_n| + |c_{-n}| = \mathcal{O}(n^{-\beta})$, then

$$\sup_{x \in [0,1]} \Big| \sum_{i=1}^{n} (f(x + \frac{i}{n}) - \int_{[0,1]} f(t)dt) \Big| = \mathcal{O}(n^{1-\beta}).$$

When the functions $f_1, \ldots, f_d$ can be expanded into a Fourier series

$$f_j(x) = \sum_{h \in \mathbb{Z}} c_h^{(j)} e^{2\pi i h x}$$

where the coefficients $(c_h^{(j)})_{h \in \mathbb{Z}}$ satisfy : $c_0^{(j)} = \frac{1}{2d}$ and there exists $\beta_j > 1$ such that

$$|c_n^{(j)}| + |c_{-n}^{(j)}| = \mathcal{O}(n^{-\beta_j}),$$

then the proof of Theorem 4.2 can be adapted so as to get a functional LDP for the $n-$dynamic $\mathbb{Z}^d$-random walks.

4.4 Dynamic linear lists

From Table 1, the evolution of dynamic linear lists is modeled by the one-dimensional dynamic random walk $(S_k)_{0 \leq k \leq n}$ each path being assigned relative weight

$$\prod_{i=1}^{n} (S_{i-1} + 1)$$

and conditioned to end in 0. We have indeed to take into account the number of places where we can delete or insert an item in the list. Let $(S^w_{ll,k})_{0\leq k\leq n}$ be the weighted random walk and

$$S^w_{ll,n}(t) := \frac{S^w_{ll,[nt]}}{n},\ t\in[0,1].$$

We use here the same notation as in Section 3 with $d=1$. We assume that the system is uniquely ergodic. The function f_1 is denoted by f and we assume

$$\int_E |\log[f(1-f)]|\ d\mu\ <\ \infty. \tag{4.2}$$

Let us recall that $\mathbb{P}_{x,n}$ denotes the distribution of the random variable $\left(\frac{S_{[nt]}}{n}\right)_{t\in[0,1]}$. For every $n\geq 1$, we define the functional $F_n:\Omega\longrightarrow[-\infty,+\infty)$ by

$$F_n(\phi(.)) = \begin{cases} \int_0^1 \log\ (\phi(t)+1/n)\ dt, \text{ if } \phi(1)=0. \\ -\infty, \text{ otherwise,} \end{cases}$$

as well as the functional

$$F(\phi(.)) = \begin{cases} \int_0^1 \log\ \phi(t)\ dt, \text{ if } \phi(1)=0 \\ -\infty, \text{ otherwise.} \end{cases}$$

With this notation, let us define a probability measure on $L_\infty([0,1])$ by

$$\mathbb{Q}^{(ll)}_{x,n}(A) = \frac{\int_A \exp(nF_n(\phi))\ d\mathbb{P}_{x,n}}{\int_\Omega \exp(nF_n(\phi))\ d\mathbb{P}_{x,n}}$$

or, in other words,

$$\mathbb{Q}^{(ll)}_{x,n}(A) = \frac{\mathbb{E}_x([\prod_{i=1}^n (S_{i-1}+1)]1_{S_n(.)\in A}\mid S_n=0)}{\mathbb{E}_x([\prod_{i=1}^n (S_{i-1}+1)]\mid S_n=0)}.$$

Let us stress that this probability forces the path to remain nonnegative.

Lemma 4.1. *The functional $I-F$ has an unique minimizer denoted by ϕ_{ll}. This function is absolutely continuous, concave and satisfies $\dot\phi_{ll}(0)>0$, $\dot\phi_{ll}(1)<0$ and $\phi_{ll}>0$ on $(0,1)$. Furthermore, assume that ϕ_{ll} has a continuous second order derivative. Let $L_{ll}(x,y)=\Lambda^*(y)-\log x$. Then the minimizer ϕ_{ll} is a solution of the Euler-Lagrange equation*

$$\frac{d}{dt}\frac{\partial L_{ll}}{\partial y}(\phi,\dot\phi) - \frac{\partial L_{ll}}{\partial x}(\phi,\dot\phi) = 0 \tag{4.3}$$

with boundary conditions $\phi(0)=\phi(1)=0$.

Proof :
Existence and unicity of the minimizer :
The existence of the minimizer is an application of Exercise 4.3.10 in [19]. We use the upper semicontinuity of F and the fact that F is bounded above on any set where the good rate function I is finite. The uniqueness of the minimizer follows from the strict convexity of the functional $I - F$ on its domain.
Properties of the minimizer ϕ_{ll} :
The domain of the functional $I - F$ is included in the set of absolutely continuous non negative functions vanishing at points 0 and 1. Hence the minimizer ϕ_{ll} must be an absolutely continuous non negative function such that $\phi_{ll}(0) = \phi_{ll}(1) = 0$. Suppose that it is not concave. Define the function $\tilde{\phi}$ by

$$\tilde{\phi}(t) = \sup\{\int_D \dot{\phi}_{ll}(t) \mid D \subset [0,1], |D| = t\}.$$

It would verify $I(\phi_{ll}) = I(\tilde{\phi})$ and $F(\phi_{ll}) < F(\tilde{\phi})$. This would contradict the fact that ϕ_{ll} is a minimizer of $I - F$. Hence, the function ϕ_{ll} is a nonzero concave function vanishing at points 0 and 1. This implies that $\dot{\phi}_{ll}(0) > 0$, $\dot{\phi}_{ll}(1) < 0$ and $\phi_{ll} > 0$ on $(0,1)$.
Euler-Lagrange equation :
The link between functional minimization and Euler-Lagrange equation is not straightforward since we work on the space of absolutely continuous functions. Euler-Lagrange equation is a second order differential equation, and we can not take for granted that ϕ_{ll} is twice differentiable. However, if we suppose that ϕ_{ll} has a continuous second order derivative, then it satisfies the Euler-Lagrange equation

$$\frac{d}{dt}\frac{\partial L_{ll}}{\partial y}(\phi, \dot{\phi}) - \frac{\partial L_{ll}}{\partial x}(\phi, \dot{\phi}) = 0 \tag{4.4}$$

where $L_{ll}(x,y) = \Lambda^*(y) - \log x$ with boundary conditions $\phi(0) = \phi(1) = 0$. This is a standard result from calculus of variations. □

Theorem 4.3. *For $x \in E$, the sequence $(S^w_{ll,n})$ converges in $\mathbb{Q}^{(ll)}_{x,n}$-probability to ϕ_{ll} as n goes to infinity and this convergence is exponential : for any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that for all sufficiently large n,*

$$\mathbb{Q}^{(ll)}_{x,n}(\{\phi \in \Omega \mid \|\phi - \phi_{ll}\|_\infty \geq \varepsilon\}) \leq \exp(-n\delta).$$

Proof :
The proof is essentially based on an adaptation in our particular case of Varadhan's integral lemma (see Section 4.3 in [4]). It will be deduced from both following inequalities. For any closed set $A \subset \Omega$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \left[\int_A \exp(nF_n)\, d\mathbb{P}_{x,n}\right] \leq - \inf_{\phi \in A} (I - F)(\phi) \tag{4.5}$$

and

$$\liminf_{n\to\infty} \frac{1}{n} \log \left[\int_\Omega \exp(nF_n)\ d\mathbb{P}_{x,n} \right] \geq - \inf_{\phi\in\Omega} (I-F)(\phi) = F(\phi_{ll}) - I(\phi_{ll}) \quad (4.6)$$

where I is given in Theorem 4.2.
Proof of (4.5) : It is a consequence of the proof of Lemma 4.3.6 of [4]. From the one hand, F_n is bounded above by 2. From the other hand, for fixed n, F_n is upper semicontinuous. The dependence of F_n on n needs a slight adaptation of the proof, which is easy. □

Proof of (4.6) : The functions F_n and F are not lower semi-continuous, so we are not able to use Lemma 4.3.4 of [19] to establish the lower bound. Let $B(\epsilon) = \{\phi : \phi > \phi_{ll} \text{ on } [\epsilon, 1-\epsilon]\}$. We have

$$\begin{aligned}
& \mathbb{E}_{x,n}\left(\exp(nF_n)\right) \\
\geq\ & \int_{B(\epsilon)} \exp(nF_n(\phi))\ d\mathbb{P}_{x,n}(\phi) \\
=\ & \int_{B(\epsilon)} \exp\left(n \int_0^1 \log\ (\phi(t)+1/n)\ dt \right)\ d\mathbb{P}_{x,n}(\phi) \\
\geq\ & \exp\left(n \int_\varepsilon^{1-\varepsilon} \log\ (\phi_{ll}(t)+1/n)\ dt \right) \\
& \times \int_{B(\epsilon)} \exp\left(n \int_0^\varepsilon \log\ (\phi(t)+1/n)\ dt \right) \exp\left(n \int_{1-\varepsilon}^1 \log\ (\phi(t)+1/n)\ dt \right)\ d\mathbb{P}_{x,n}(\phi)
\end{aligned}$$

In the last integral, let us condition on $\phi(\varepsilon)$ and $\phi(1-\varepsilon)$ and apply the Markov property. This yields

$$\begin{aligned}
& \int_{B(\epsilon)} \exp\left(n \int_0^\varepsilon \log\ (\phi(t)+1/n)\ dt \right) \exp\left(n \int_{1-\varepsilon}^1 \log\ (\phi(t)+1/n)\ dt \right)\ d\mathbb{P}_{x,n}(\phi) \\
=\ & \mathbb{E}_{x,n}[\ Y_1\ Y_2\ Y_3\ ].
\end{aligned}$$

where

$$\begin{aligned}
Y_1 &= \mathbb{P}_{x,n}[B(\varepsilon) \mid \phi(\varepsilon),\ \phi(1-\varepsilon)\ ], \\
Y_2 &= \mathbb{E}_{x,n}[\ \exp(n \int_0^\varepsilon \log\ (\phi(t)+1/n)\ dt) \mid \phi(\varepsilon)\ ], \\
Y_3 &= \mathbb{E}_{x,n}[\ \exp(n \int_{1-\varepsilon}^1 \log\ (\phi(t)+1/n)\ dt) \mid \phi(1-\varepsilon)\ ].
\end{aligned}$$

If $\phi(\varepsilon) \leq \phi_{ll}(\varepsilon)$ or $\phi(1-\varepsilon) \leq \phi_{ll}(1-\varepsilon)$, then $Y_1 = 0$.
Let us define

$$K_n^\varepsilon = \inf_{y \geq \phi_{ll}(\varepsilon)} \mathbb{E}_{x,n}[\ \exp(n \int_0^\varepsilon \log\ (\phi(t)+1/n)\ dt) \mid \phi(\varepsilon) = y\ ]$$

and

$$L_n^\varepsilon = \inf_{y \geq \phi_{ll}(1-\varepsilon)} \mathbb{E}_{x,n}[\ \exp(n \int_{1-\varepsilon}^1 \log\ (\phi(t) + 1/n)\ dt) \mid \phi(1-\varepsilon) = y\ ].$$

Then, $Y_1\ Y_2\ Y_3\ \geq Y_1\ K_n^\varepsilon\ L_n^\varepsilon$ and

$$\mathbb{E}_{x,n}[\ Y_1\ Y_2\ Y_3\ ] \geq \mathbb{E}_{x,n}[\ Y_1\ ]\ K_n^\varepsilon\ L_n^\varepsilon = \mathbb{P}_{x,n}(B(\varepsilon))\ K_n^\varepsilon\ L_n^\varepsilon.$$

This yields

$$\begin{aligned} &\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{x,n}(\exp(nF_n)) \\ \geq\ & \int_\varepsilon^{1-\varepsilon} \log\ \phi_{ll}(t)\ dt + \liminf_{n\to\infty} \frac{1}{n} \log\ \mathbb{P}_{x,n}(B(\varepsilon)) + \liminf_{n\to\infty} \frac{1}{n} \log K_n^\varepsilon + \liminf_{n\to\infty} \frac{1}{n} \log L_n^\varepsilon. \end{aligned}$$

The set $B(\varepsilon)$ is open, so from the lower bound of large deviations (Theorem 3.2), we get :

$$\liminf_{n\to\infty} \frac{1}{n} \log\ \mathbb{P}_{x,n}(B(\varepsilon)) \geq -\inf\{I(\phi); \phi \in B(\varepsilon)\}.$$

Now, thanks to the regularity of I,

$$\inf\{I(\phi); \phi \in B(\varepsilon)\} = \inf\{I(\phi); \phi \in \bar{B}(\varepsilon)\} \leq I(\phi_{ll}).$$

Letting $\varepsilon \to 0$, this yields

$$\begin{aligned} &\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{x,n}(\exp(nF_n)) \\ \geq\ & F(\phi_{ll}) - I(\phi_{ll}) + \limsup_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log K_n^\varepsilon + \limsup_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log L_n^\varepsilon. \end{aligned}$$

Let $y > \phi_{ll}(\varepsilon) \neq 0$. The conditional expectation

$$\mathbb{E}_{x,n}[\ \exp(n \int_0^\varepsilon \log\ (\phi(t) + 1/n)\ dt) \mid \phi(\varepsilon) = y\ ]$$

is underestimated by the contribution of the path having increments $+1$ between times 0 and $([n\varepsilon]+ny)/2$ and increments -1 between times $([n\varepsilon]+ny)/2$ and $[n\varepsilon]$. For this path, $\phi(t) + 1/n > \phi_{ll}(\varepsilon)t$ on $[0, \varepsilon]$, hence

$$\begin{aligned} &\mathbb{E}_{x,n}[\ \exp(n \int_0^\varepsilon \log\ (\phi(t) + 1/n)\ dt) \mid \phi(\varepsilon) = y\ ] \\ \geq\ & \exp(n \int_0^\varepsilon \log\ (\phi_{ll}(\varepsilon)t)\ dt\ ) \prod_{i=1}^{([n\varepsilon]+ny)/2} f(T^i x) \prod_{i=([n\varepsilon]+ny)/2+1}^{[n\varepsilon]} (1 - f(T^i x)) \end{aligned}$$

Thus,

$$\frac{1}{n} \log K_n^\varepsilon \geq \int_0^\varepsilon \log\ (\phi_{ll}(\varepsilon)t)\ dt + \frac{1}{n} \sum_{i=1}^{[n\varepsilon]} \log f(T^i x) + \frac{1}{n} \sum_{i=1}^{[n\varepsilon]} \log\ (1 - f(T^i x)).$$

Then, the uniform ergodicity of the dynamical system and hypothesis (4.2) imply that

$$\liminf_{n\to\infty} \frac{1}{n} \log K_n^\varepsilon \geq \int_0^\varepsilon \log\ (\phi_{ll}(\varepsilon)t)\ dt + \varepsilon \int_E \log\ [f(1-f)]\ d\mu.$$

Finally, since $\phi_{ll}(0) = 0$ and $\dot{\phi_{ll}}(0) > 0$ (see Lemma 4.1),

$$\limsup_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log(K_n^\varepsilon) = 0.$$

In the same way, we prove that

$$\limsup_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log(L_n^\varepsilon) = 0.$$

This proves inequality (4.6).
Let us end the proof of Theorem 4.3. Inequalities (4.5) and (4.6) imply that $\mathbb{Q}_{x,n}^{(ll)}$ satisfies a large deviation upper bound : for every closed set $A \subset \Omega$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{Q}_{x,n}^{(ll)}(A) \leq - \inf_{\phi\in A} \left[\ I(\phi) - F(\phi) - \inf_{\phi\in\Omega}(\ I(\phi) - F(\phi)\ )\right].$$

Apply this inequality with $A_\varepsilon = \{\ \phi \in \Omega \mid \|\phi - \phi_{ll}\|_\infty \geq \varepsilon\}$. The function F is upper semi-continuous. On the set where the good rate function I is finite, the function F is bounded above (because $I(\phi) < +\infty$ implies $|\dot{\phi}| \leq 1$). Thus, by applying the result of Exercise 4.3.10 in [19], we deduce that for any closed set $C \subseteq \Omega$, the infimum of $I - F$ on the set C is attained. Since ϕ_{ll} does not belong to A_ε, the infimum over A_ε is positive. So there exists $\delta = \delta(\varepsilon) > 0$ such that for all sufficiently large n,

$$\mathbb{Q}_{x,n}^{(ll)}(A_\varepsilon) \leq \exp(-n\delta).$$

□

Remark :
Routine calculations give, for every λ

$$\Lambda'(\lambda) = \int_E \frac{fe^\lambda - (1-f)e^{-\lambda}}{fe^\lambda + (1-f)e^{-\lambda}}\, d\mu,$$

$$\Lambda''(\lambda) = 4 \int_E \frac{f(1-f)}{[fe^\lambda + (1-f)e^{-\lambda}]^2}\, d\mu.$$

Under assumption (4.2), we have $\int_E f(1-f)\ d\mu > 0$ so that $\Lambda'' > 0$ and Λ' is an homeomorphism from $\mathbb{R}$ to $(-1, 1)$. Moreover

$$(\Lambda^*)'' = ((\Lambda')^{-1})'.$$

Equation (4.3) is then equivalent to

$$\ddot{\phi}(t)((\Lambda')^{-1})'(\dot{\phi}(t)) = -\frac{1}{\phi(t)} \tag{4.7}$$

with boundary conditions $\phi(0) = \phi(1) = 0$.
When $f \equiv 1/2$, we have $\Lambda(\lambda) = \log\cosh\lambda$, hence the unique solution of (4.7) with boundary conditions : $\phi(0) = \phi(1) = 0$ is given by

$$\phi_{ll}(t) = \frac{1}{\pi}\sin(\pi t)$$

(see [41] or [44] for further details). This example is the simplest one. For other particular functions f, a solution of the above functional equation can perhaps be obtained with numerical methods. This question is hard and is actually under consideration. We will give in Section 7 an example of function f where direct calculations lead to a degenerate non linear partial differential equation. Consider the storage cost function

$$C_{ll,n} = n\sum_{i=1}^{n} S^{w}_{ll,n}(\frac{i}{n}).$$

The next result is easily derived from Theorem 4.3.

Corollary 4.1. *Under the hypothesis* (4.2)*, for any $x \in E$, the random variables $\left(\frac{C_{ll,n}}{n^2}\right)_{n\geq 1}$ converge exponentially fast to*

$$m_{ll} = \int_0^1 \phi_{ll}(t)\ dt$$

as n goes to infinity.

Remark :
When $f \equiv 1/2$, under the assumption that ϕ_{ll} is $\mathcal{C}^2$, then $\phi_{ll}(t) = \frac{1}{\pi}\sin(\pi t)$ and $m_{ll} = \frac{2}{\pi^2}$.

4.5 Dynamic priority queues

This section deals more briefly with priority queues driven by the dynamic random walk defined in Section 3. From Table 1, we see that the difference between dynamic priority queues and dynamic linear lists is the weight we assign to each path. These queues are modeled by the one-dimensional dynamic random walk (S_k) each path being assigned relative weight, and conditioned on $S_n = 0$. It comes from the fact that in the priority queue case, the number

of insertions is equal to the number of deletions ; the structure beginning and ending empty. We shall denote this weighted random walk by $(S^w_{pq,n})_{0\le k\le n}$. The normalized data structure size as a function of time is defined by

$$S^w_{pq,n}(t) := \frac{S^w_{pq,[nt]}}{n}, \quad t \in [0,1].$$

The distribution of the random variable $(S^w_{pq,n}(t))_{t\in[0,1]}$ with values in Ω is denoted by $\mathbb{Q}^{(pq)}_{x,n}$.
The situation is similar to Section 4, with instead of L_{ll} the function

$$L_{pq}(x,y) = \Lambda^*(y) - \frac{1}{2}\log x,$$

so that the associated Euler-Lagrange equation is

$$\ddot{\phi}(t)(\Lambda^*)''(\dot{\phi}(t)) = -\frac{1}{2\phi(t)},$$

whose solution is denoted by ϕ_{pq}. The storage cost is

$$C_{pq,n} = n\sum_{i=1}^{n} S^w_{pq,n}(\frac{i}{n}).$$

We have again exponential convergence, of $S^w_{pq,n}$ to ϕ_{pq} and of $\frac{C_{pq,n}}{n^2}$ to

$$m_{pq} = \int_0^1 \phi_{pq}(t)\ dt.$$

Remark :
When $f \equiv 1/2$, under the assumption that ϕ_{pq} is $\mathcal{C}^2$, then $\phi_{pq}(t) = t(1-t)$ and $m_{pq} = \frac{1}{6}$.

4.6 Dynamic dictionaries

4.6.1 A new dynamic random walk

Because of the operations supported by dictionaries, we need a different model of dynamic random walks.
We keep the notation of the previous section, except that for each $x \in E$ and $i \ge 1$ the law of X_i is

$$\mathbb{P}_x(X_i = z) = \begin{cases} \frac{1}{2}f(T^i x) & \text{if } z = 1 \\ \frac{1}{2}f(T^i x) & \text{if } z = -1 \\ 1 - f(T^i x) & \text{if } z = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then the same results as above hold, with

$$\Lambda_d(\lambda) = \int_E \log\Big(1 + (\cosh(\lambda) - 1)f\Big)d\mu.$$

4.6.2 Large deviation principles

All results of Section 3.2 (Theorems 3.1 and 3.2) hold with, instead of Λ,

$$\Lambda_d(\lambda) = \mu\Big(\log\Big(1 + (\cosh(\lambda) - 1)f\Big) \;\Big|\; \mathcal{I}\Big).$$

When the system is uniquely ergodic, we get

$$\Lambda_d(\lambda) = \mu\Big(\log\Big(1 + (\cosh(\lambda) - 1)f\Big)\Big).$$

In the same way, Theorem 4.1 and Corollary 4.1 hold, up to a change of notation. The Euler-Lagrange equation is now :

$$\ddot{\phi}(t)(\Lambda_d^*)''(\dot{\phi}(t)) = -\frac{1}{\phi(t)} \tag{4.8}$$

where

$$\Lambda_d^*(y) = \sup_{\lambda}\{\lambda\, y - \Lambda_d(\lambda)\}.$$

Remark :
When $f \equiv 1/2$, we have $\Lambda_d^* = 2\Lambda^*$ (see [41] or [44]) and thus, under the assumption that ϕ_d is $\mathcal{C}^2$, the solution ϕ_d of equation (4.8) is the same as in the case of priority queues, namely

$$\phi_d(t) = t(1-t), t \in [0,1]$$

and $m_{pq} = m_d = \frac{1}{6}$.

4.7 An example : Linear lists and rotation on the torus

Let $([0,1], \mathcal{B}([0,1]), \lambda, T_\alpha)$ be the dynamical system where T_α is defined by $x \to x + \alpha \mod 1$, with α a given real and λ is the Lebesgue measure on $[0,1]$. This particular dynamical system is the so-called rotation on the one-dimensional torus. Twofold motivations are related to this example :
1. Explicit calculations are possible in this case,
2. When α is rational, we get a periodic dynamical system which models a periodic behaviour of the operations on the data structures.
Irrational rotations are uniquely ergodic, i.e. the ergodic average of a continuous function uniformly converges in $x \in [0,1]$ to the integral of f. The uniform convergence of the ergodic averages even holds for any function with bounded variation (see [36]). Consequently, if we consider a dynamic random walk associated to an irrational rotation on the torus, Theorem 4.1 holds for every function f with bounded variation.

4.7.1 Choice of a function and derivation of the corresponding differential equation

When the hypothesis (4.2) is not satisfied by the function f, the methods presented in the paper do not apply. For instance, let us choose the function $f = 1_{[0,\frac{1}{2}]}$, then $\Lambda = 0$ and

$$\Lambda^*(y) = \begin{cases} +\infty \text{ if } y \neq 0 \\ 0 \quad \text{otherwise.} \end{cases}$$

In that case, no large deviation occurs as the dynamic random walk is deterministic in the sense that

$$S_n = \sum_{i=1}^{n} (21_{[0,\frac{1}{2}]}(T_\alpha^i x) - 1)$$

for some $x \in [0,1]$ fixed.

The computation of the function Λ^* for a general function f is difficult (see the remark following Theorem 4.1). However, we are able to compute it for the very particular function $f = \frac{3}{4}1_{[0,\frac{2}{3}]}$. Unfortunately, this function does not satisfy hypothesis (4.2). We think that the hypothesis (4.2) is only technical and should be dropped out but we don't have any proof of this point yet. In order to illustrate how our method is efficient to determine the asymptotic behavior of the storage cost function associated to a dynamic linear list, we now compute the function ϕ_{ll} for this particular function f, under the assumption that ϕ_{ll} is $\mathcal{C}^2$. Straightforward computations give us

$$\Lambda(\lambda) = \log(\cosh(\lambda)) + \frac{2}{3}\log\left(1 + \frac{\tanh(\lambda)}{2}\right) + \frac{1}{3}\log\left(1 - \tanh(\lambda)\right).$$

When $y \in]-1, \frac{1}{3}[$,

$$\Lambda^*(y) = \frac{1}{2}(1+y)\log(1+y) + \frac{(1-3y)}{6}\log(1-3y)$$

and $+\infty$ otherwise.
We consider dynamic linear lists driven by this particular dynamic random walk and determine the path ϕ_{ll} satisfying the following Euler-Lagrange equation :

$$\ddot{\phi}(t)(\Lambda^*)''(\dot{\phi}(t)) = -\frac{1}{\phi(t)}. \tag{4.9}$$

In this particular case, on the interval $]-1, 1/3[$,

$$(\Lambda^*)''(y) = ((\Lambda')^{-1})'(y) = \frac{1}{2(1+y)} + \frac{3}{2(1-3y)}.$$

After straightforward computations, the equation (4.9) becomes

$$2\phi\ddot{\phi} = 3\dot{\phi}^2 + 2\dot{\phi} - 1. \tag{4.10}$$

4.7.2 Study of the differential equation (4.10)

Rewriting the equation like a system of two differential equations, with $\theta = \dot{\phi}$ gives :

$$\begin{cases} \dot{\phi} = \theta \\ \dot{\theta} = \frac{3}{2\phi}(\theta - \frac{1}{3})(\theta + 1) \end{cases} \iff \dot{z} = F(z) \text{ with } z(t) = \begin{bmatrix} \phi(t) \\ \theta(t) \end{bmatrix}$$

which shows that a solution orbit $(\phi(t), \theta(t))$ may cross the $\phi = 0$ line only when $\theta = \frac{1}{3}$ or $\theta = -1$. We note also that the curves $(\phi(t) = \frac{1}{3}(t - t_0) + \phi_0, \theta(t) = \frac{1}{3})$ and $(\phi(t) = -(t - t_0) + \phi_0, \theta(t) = -1)$ are solutions of the equation, moreover they are the only polynomial solutions.

A first overview of the behavior of the differential equation may be suggested by plotting the (normalized) vector field F on a grid in the phase space, see Figure 4.2 (on the figure the letters r,l stand for "right" and "left" and the letters u,m and d for "upper", "middle" and "down" ; also x and y are used in place of (respectively) ϕ and θ).

FIG. 4.2 – Normalized vector field and symmetry of trajectories

The vector field suggests a symmetry of the trajectories (see Figure 4.2). Furthermore we have also a similarity principle between some solution orbits. More precisely we have :

Lemma 4.2. *if* $(\phi(t), \theta(t))$, $t \in [0, T]$ *is a trajectory of the differential equation then :*

– *(symmetry)* $(\hat{\phi}(t), \hat{\theta}(t))$, $t \in [c, c+T]$ *defined by :*

$$\begin{cases} \hat{\phi}(t) = -\phi(c+T-t) \\ \hat{\theta}(t) = \theta(c+T-t)) \end{cases}$$

– *and (similarity)* $(\tilde{\phi}(\tau), \tilde{\theta}(\tau))$, $\tau \in [c, c+T/\alpha]$, *with* $\tau = t/\alpha + c$ *defined by :*

$$\begin{cases} \tilde{\phi}(\tau) = \frac{1}{\alpha}\phi(t) \\ \tilde{\theta}(\tau) = \theta(t) \end{cases}$$

where $\alpha > 0$ *and* c *are two constants, are also two trajectories of the differential equation.*

Proof :
Straightforward computations show that $(\hat{\phi}(t), \hat{\theta}(t)), \forall t \in [c, c+T]$ and $(\tilde{\phi}(\tau), \tilde{\theta}(\tau))$, $\forall \tau \in [c, c+T/\alpha]$ verify the differential equation since $(\phi(t), \theta(t))$, $\forall t \in [0, T]$ verifies it. □

Thanks to the symmetry property, it is enough to study the dynamic on the half plane $\phi \geq 0$. This property (together with the similarity property) will also be very useful for the computation of solutions : it will be enough to compute accurately very few trajectories (in fact 3) to get all the others.

Proposition 4.3. *Starting from the initial condition* (ϕ_0, θ_0), *with* $\phi_0 > 0$, *the dynamical behavior may be summarized as follows :*
– *if* $(\phi_0, \theta_0) \in E_{r,u} = \{\phi > 0\} \times \{\frac{1}{3} < \theta\}$, *then the trajectory stays in the set* $E_{r,u}$, $\phi(t)$ *and* $\theta(t)$ *are increasing on* $[0, +\infty)$ *and :*

$$\lim_{t \to +\infty} \phi(t) = \lim_{t \to +\infty} \theta(t) = +\infty.$$

– *if* $(\phi_0, \theta_0) \in E_{r,m} = \{\phi > 0\} \times \{-1 < \theta < \frac{1}{3}\}$, *then the trajectory stays in the set* $E_{r,m}$ *until a time* $\bar{t}$ *where* $(\phi(\bar{t}), \theta(\bar{t})) = (0, -1)$ *with :*

$$\bar{t} = \int_0^{+\infty} 2\phi_0 e^{\frac{2\xi}{3}} \left| \frac{C-1}{Ce^{4\xi}-1} \right|^{\frac{2}{3}} d\xi, \quad \text{and } C = C(\theta_0) = \frac{\theta_0 - \frac{1}{3}}{\theta_0 + 1}$$

Moreover $\theta(t)$ *is decreasing on* $[0, \bar{t}]$ *and starting from* $\theta_0 > 0$, $\phi(t)$ *first increases until a time* $\hat{t}$ *(which corresponds to* $\phi(\hat{t}) = 0$*) where* $\phi(t)$ *becomes to decrease.*
– *if* $(\phi_0, \theta_0) \in E_{r,d} = \{\phi > 0\} \times \{\theta < -1\}$, *then the trajectory stays in the set* $E_{r,d}$ *until the time* $\bar{t}$ *where* $(\phi(\bar{t}), \theta(\bar{t})) = (0, -1)$.
– *the points* $(0, \frac{1}{3})$ *and* $(0, 1)$ *are singular for the dynamic : there are an infinity number of trajectories across them. Any trajectory starting from* (ϕ_0, θ_0) *in* $E_{r,m}$, $E_{r,d}$, $E_{l,m}$, *or* $E_{l,u}$ *reaches one of the 2 singular points but may be extended (after the singularity) in the previous mentioned symmetric way (which lets to have the maximum regularity for these completed trajectories). In particular a trajectory starting at* $(\phi_0, \theta_0) \in E_{r,m}$ *may be extended at time*

$\bar{t}$ in $E_{l,m}$ (by the $x = 0$ axis symmetry). Arriving at $(0, \frac{1}{3})$ this second part of the trajectory may also be extended in $E_{r,m}$ and finally reaches the point (ϕ_0, θ_0) at time T (depending on (ϕ_0, θ_0)) giving rise to a periodic trajectory. For these periodic orbits, which cut the $\theta = 0$ axis, say for $\phi = \Phi$, the period T is a linear function of Φ, more precisely :

$$T = \frac{2^{4/3}}{\sqrt{3}}\beta(\frac{1}{6}, \frac{1}{2})\Phi \simeq 10.6\Phi.$$

– *The integral curves are given by :*

$$\phi = \phi_0 \left|\frac{\theta - 1/3}{\theta_0 - 1/3}\right|^{\frac{1}{6}} \left|\frac{\theta + 1}{\theta_0 + 1}\right|^{\frac{1}{2}} \tag{4.11}$$

Proof :
The main tool consists in applying a time-change $(t \mapsto \tau)$ defined by :

$$\begin{cases} \frac{d\tau}{dt} = \frac{1}{2\phi} \\ \tau(t = 0) = 0 \end{cases} \quad \text{while } \phi > 0,$$

because we can get the trajectories analytically as function of τ. The function θ verifies the differential equation :

$$\frac{d\theta}{d\tau} = 3(\theta - \frac{1}{3})(\theta + 1), \text{ while } \phi > 0,$$

which yields

$$\theta(\tau) = \frac{\frac{1}{3} + Ce^{4\tau}}{1 - Ce^{4\tau}}, \text{ while } \phi > 0 \tag{4.12}$$

with :

$$C = C(\theta_0) = \frac{\theta_0 - \frac{1}{3}}{\theta_0 + 1} \text{ and } \begin{cases} C > 1 \text{ when } \theta_0 \in (-\infty, -1) \\ C < 0 \text{ when } \theta_0 \in (-1, \frac{1}{3}) \\ C \in (0, 1) \text{ when } \theta_0 \in (\frac{1}{3}, +\infty) \end{cases}$$

Starting from $(\phi_0, \theta_0) \in E_{r,u}$, we have :

$$\lim_{\tau \to \bar{\tau}-} \theta(\tau) = +\infty, \ \bar{\tau} = -\log(C)/4.$$

We will see later on that in this case the time t goes also to $+\infty$ (as a function of τ) and so there is no blow up at some finite time $\bar{t}$.
The expression of ϕ in terms of τ can be obtained via :

$$\frac{d\phi}{d\tau} = \frac{d\phi}{dt}\frac{dt}{d\tau} = 2\phi\theta, \text{ while } \phi > 0.$$

We get after some computations :

$$\phi(\tau) = \phi_0 e^{\frac{2\tau}{3}} \left|\frac{C - 1}{Ce^{4\tau} - 1}\right|^{\frac{2}{3}}, \text{ while } \phi > 0. \tag{4.13}$$

Like for $\theta(\tau)$ we see that, starting from $(\phi_0, \theta_0) \in E_{r,u}$ we have :

$$\lim_{\tau \to \bar{\tau}-} \phi(\tau) = +\infty,\ \bar{\tau} = -\log(C)/4.$$

Finally, from $dt = 2\phi d\tau$ we have the relation giving the time t in term of τ :

$$t(\tau) = \int_0^{\tau} 2\phi_0 e^{\frac{2\xi}{3}} \left| \frac{C-1}{Ce^{4\xi}-1} \right|^{\frac{2}{3}} d\xi,\ \text{while } \phi > 0. \tag{4.14}$$

We can now prove the stated behavior :

– if $(\phi_0, \theta_0) \in E_{r,u}$ then the time-change $t \mapsto \tau$ is a diffeomorphism from $[0, +\infty)$ to $[0, \bar{\tau})$ ($\bar{\tau} = \log(C)/4$). Combined with (4.12) and (4.13) we have then :

$$\lim_{t \to +\infty} \phi(t) = \lim_{t \to +\infty} \theta(t) = +\infty,$$

– if $(\phi_0, \theta_0) \in E_{r,m}$, we get $C < 0$ and the time-change is a diffeomorphism from $[0, \bar{t})$ to $[0, +\infty)$ with :

$$\bar{t} = \int_0^{+\infty} 2\phi_0 e^{\frac{2\xi}{3}} \left| \frac{C-1}{Ce^{4\xi}-1} \right|^{\frac{2}{3}} d\xi$$

and (4.12) and (4.13) show that :

$$\lim_{t \to \bar{t}-} \theta(t) = -1,\ \lim_{t \to \bar{t}-} \phi(t) = 0$$

Furthermore, computing derivatives, we see that θ is strictly decreasing on $[0, \bar{t})$ and :

– if $\theta_0 > 0$, then ϕ is increasing on $[0, t(\hat{\tau}))$, and decreasing on $(t(\hat{\tau}), \bar{t}))$, with :

$$\hat{\tau} = -\frac{1}{4}\log(-3C)$$

corresponding to the time when $\theta = 0$.

– if $\theta_0 \leq 0$, $\phi(t)$ is decreasing on $[0, \bar{t})$.

It is obvious that a trajectory starting from $(\phi_0, \theta_0) \in E_{r,u}$ and reaching $(0, -1)$ at time $\bar{t}$, may be completed by a first part connecting $(0, 1/3)$ to (ϕ_0, θ_0) in time $-t(-\infty)$. The family of all these trajectories going from $(0, 1/3)$ to $(0, -1)$ may be parametrized in a unique way by $\Phi > 0$, Φ selecting the only one passing at the point $(\Phi, 0)$. The time to go from $(0, 1/3)$ to $(0, -1)$ is :

$$\int_{-\infty}^{+\infty} 2\phi_0 e^{\frac{2\xi}{3}} \left| \frac{C-1}{Ce^{4\xi}-1} \right|^{\frac{2}{3}} d\xi = \frac{2^{1/3}}{\sqrt{3}} \beta(\frac{1}{6}, \frac{1}{2})\Phi \tag{4.15}$$

If we complete such a trajectory by the symmetrized trajectory in $E_{l,m}$ we have then a periodic orbit of period $T = 2^{4/3}/\sqrt{3}\ \beta(1/6, 1/2)\Phi$.

– if $(\phi_0, \theta_0) \in E_{r,d}$, we have $C > 1$, the time-change $t \mapsto \tau$ is a diffeomorphism from $[0, \bar{t})$ to $[0, +\infty)$. Here ϕ is decreasing while θ is increasing.

– From (4.12) and (4.13) we get (4.11), giving the integral curves independently of t.
□

Like for all analytical expressions, we can compute safely the trajectories in terms of t. In fact the only numerical work involved consists in approximating the integral (4.14), which is easy. By the way, we have also implementeded a (good) numerical differential equation solver (at the very beginning of this study) and we note that different results may be obtained when crossing (near) the singular points, where the solver may continue the trajectory to various ones. This is not surprising !

Figure (4.3) shows two periodic trajectories, one as unbroken line passing at $(1, 0)$ the other as dotted line passing at $(0.5, 0)$ and computed thanks to the similarity property.

FIG. 4.3 – Two trajectories of the differential equation

Finally we are interested in a solution $\phi_{ll}(t)$, $t \in [0, 1]$ of the differential equation (4.10) such that $\phi_{ll}(0) = \phi_{ll}(1) = 0$. If we constrain ϕ_{ll} to be positive ($t \in (0, 1)$), then this solution is unique and corresponds to the trajectory in $E_{r,m}$ which goes from $(0, 1/3)$ to $(0, -1)$ in a unit time, and so, the one

parametrized (see Equation (4.15)) by :

$$\Phi = \frac{\sqrt{3}}{2^{1/3}\beta(\frac{1}{6},\frac{1}{2})}.$$

Like the period T, the area (i.e. m_{ll}) can be computed in term of the beta function :

$$m_{ll} = \int_0^1 \phi_{ll}(t)dt = \int_{-\infty}^{+\infty} \phi_{ll}(t(\tau))\frac{dt}{d\tau}\,d\tau = \frac{6}{\beta(\frac{1}{6},\frac{1}{2})^2} \simeq 0.113.$$

FIG. 4.4 – The positive solution such that $\phi_{ll}(0) = \phi_{ll}(1) = 0$

4.8 Concluding remarks

We have shown that is is possible to analyze dynamic data structures when the distribution on the set of histories is not uniform and when the operations are modeled by time dependent dynamic random variables. We have recovered some results of Louchard (see [41]) in a more general setting but the temporal inhomogeneity of the dynamic random walk does not allow us to establish a precise large deviation principle from which we could derive the asymptotic distributions of data structure cost functions. Since our stochastic model is very general, it can be applied to a variety of real-world phenomena (including parallel and distributed computing (see [32]), modeling of multi-agents behaviors and option pricing in financial markets, ...).

Extensions et questions ouvertes

Vers des modélisations plus complexes de la strucure de l'ADN

Nous avons étudié un modèle particulier pour la dénaturation de l'ADN, relativement simple. Dans les modèles plus évolués, l'hétérogénéité du champ est souvent un obstacle à des résultats théoriques précis. Il serait intéressant de considérer ces modèles, et de voir si nos hypothèses de convergence des champs sont pertinentes pour déterminer le comportement asymptotique des systèmes.

Un TCL pour la marche pondérée ?

Dans le chapitre 3, nous avons introduit le modèle de la marche simple pondérée et étudié son asymptotique pour les fonctions fitness puissance $f(x) = x^\beta$, $\beta > 0$. Nous avons obtenu une loi des grands nombres dans l'espace des fonctions càd-làg : la marche renormalisée $\frac{1}{n} S_{[nt]}$ converge vers la fonction déterministe $\psi_\beta(t)$.

La question des fluctuations de la marche pondérée autour de la limite déterministe se pose naturellement. Nous nous attendons à un théorème du type théorème de Donsker : le processus renormalisé

$$\sqrt{n}\left(\frac{1}{n} S_{[nt]} - \psi_\beta(t)\right)$$

converge-t-il vers un processus gaussien ?

Louchard s'est intéressé à cette question [41]. Il donne la covariance du processus limite pour un modèle très proche du nôtre. Cependant, nous ne pensons pas que la preuve est faite avec toute la rigueur mathématique nécessaire. Mazza et Piau [46] prouvent un théorème central limite pour la valeur finale de la marche et pour $\beta = 1$: sous $\mathbb{Q}_n^1$, on a la convergence en loi

$$\frac{1}{\sqrt{n}}\left(S_n - \frac{2}{\pi} n\right) \Rightarrow \mathcal{N}\left(0, \frac{4}{\pi^2}\right).$$

La preuve est basée sur un calcul de la transformée de Laplace de S_n. Cette technique n'a pas pu être appliquée avec succès pour $S_{[nt]}$, et ne semble pas adaptée au cas $\beta \neq 1$.

Comment traiter le cas d'autres fonctions fitness ?

Une question naturelle est celle du comportement de la marche pondérée et de l'algorithme de sélection-mutation pour d'autres fonctions fitness. Notre étude a permis de traiter le cas des fonctions puissance, $f(x) = x^\beta$, $\beta > 0$. Quel est le comportement asymptotique de la marche pour d'autres fonctions fitness, par exemple pour un fitness logarithmique $f(x) = \log(x)$?

Pour une fonction fitness constant $f(x) = 1$, la marche a un comportement diffusif (S_n est de l'ordre $\sqrt{n}$). Pour une fonction fitness puissance $f(x) = x^\beta$, nous avons montré que la marche pondérée a un comportement ballistique avec une vitesse connue v_β. N'importe quelle vitesse peut être atteinte avec une fonction fitness appropriée puisque v_β décrit l'intervalle $(0, 1)$ lorsque β décrit $(0, +\infty)$. On peut se poser la question des régimes intermédiaires : existe-t-il une fonction fitness telle que la marche pondérée ait un comportement du type $\sqrt{n} \ll S_n \ll n$, par exemple $S_n = \Theta(n^\alpha)$ avec $1/2 < \alpha < 1$?

Une asymptotique différente pour l'algorithme génétique ?

Il s'agit ici d'étudier l'influence de la taille de la population pour les algorithmes de sélection-mutation. Considérons l'exemple que nous avons développé : l'espace d'état est $E = \mathbb{Z}$, la mutation correspond à un pas de la marche simple symétrique et la fonction fitness avec laquelle opère la sélection est $f(x) = x$.

Pour l'algorithme en population infinie, nous avons montré que le fitness croit linéairement avec le nombre d'itérations n et est équivalent asymptotiquement à $\frac{2}{\pi}n$. Le modèle en population finie a été étudie par J.Bérard et A.Bienvenüe qui ont prouvé un principe d'invariance [5]. Il découle de leur travaux que l'espérance du fitness moyen avec une population de taille p fixée et après n itérations est asymptotiquement équivalente à $(2p-1)\sqrt{n}$.

Afin de mieux comprendre le changement de régime dans cette transition population finie / population infinie, il serait intéressant d'étudier une asymptotique où la taille de la population et le nombre d'itérations tendent simultanément vers l'infini, par exemple $p_n = n$. Ces régimes intermédiaires ne sont pour l'instant pas compris et les simulations numériques ne permettent pas de conjecturer un comportement précis.

Extensions pour les structures de données dynamiques

Il serait intéressant d'étendre d'autres résultats obtenus pour les structures de données markoviennes aux structures de données dynamiques. Les techniques de modélisation en terme de marche dynamique sont suffisamment générales pour s'appliquer à d'autres types de données. Par exemple, Louchard, Kenyon, et Schott [42] ont étudié le comportement asymptotique de la taille maximale d'autres types de structures de données faisant intervenir d'autres opérations appelées "lazy deletion" et "batched insertion". Que donnerait un modèle dynamique pour ces structures de données ?

[]

Bibliographie

[1] R.J. Baxter. *Exactly solved models in statistical mechanics*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1989. Reprint of the 1982 original.

[2] C.J. Benham. Torsional stress and local denaturation in supercoiled DNA. *Proc. Natl. Sci.*, 76(8) :3870–3874, 1979.

[3] C.J. Benham. Theoretical analysis of heteropolymeric transitions in superhelical DNA at high temperature. *J. Chem. Phys.*, 92(10) :6294–6305, 1990.

[4] C.J. Benham. Theoretical of the helix-coil transition in positively superhelical DNA at high temperature. *Phys. Rev. E*, 53(3) :2984–2987, 1996.

[5] J. Bérard and A. Bienvenüe. Convergence of a genetic algorithm with finite population. In *Mathematics and computer science (Versailles, 2000)*, Trends Math., pages 155–163. Birkhäuser, Basel, 2000.

[6] J. Bérard and A. Bienvenüe. Sharp asymptotic results for simplified mutation-selection algorithms. *Ann. Appl. Probab.*, 13(4) :1534–1568, 2003.

[7] G. Bernardi. The isochore organisation of the human genome. *Annual Review of Genetics*, 1989.

[8] P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, second edition, 1986.

[9] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics : Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999.

[10] C. Boldrighini, R. A. Minlos, and A. Pellegrinotti. Almost-sure central limit theorem for a Markov model of random walk in dynamical random environment. *Probab. Theory Related Fields*, 109(2) :245–273, 1997.

[11] R. Cerf. The dynamics of mutation-selection algorithms with large population sizes. *Ann. Inst. H. Poincaré Probab. Statist.*, 32(4) :455–508, 1996.

[12] R. Cerf. Asymptotic convergence of genetic algorithms. *Adv. in Appl. Probab.*, 30(2) :521–550, 1998.

[13] O. Clay and G Bernardi. Isochores : dream or reality ? *Trends in biotechnology*, 2002.

[14] P. Clote and R. Backofen. *Computational molecular biology*. Wiley Series in Mathematical and Computational Biology. John Wiley & Sons Ltd., Chichester, 2000.

[15] F. Comets, N. Gantert, and O. Zeitouni. Quenched, annealed and functional large deviations for one-dimensional random walk in random environment. *Probab. Theory Related Fields*, 118(1) :65–114, 2000.

[16] P. Del Moral and A. Guionnet. Large deviations for interacting particle systems : applications to non-linear filtering. *Stochastic Process. Appl.*, 78(1) :69–95, 1998.

[17] P. Del Moral and A. Guionnet. On the stability of measure valued processes with applications to filtering. *C. R. Acad. Sci. Paris Sér. I Math.*, 329(5) :429–434, 1999.

[18] P. Del Moral and A. Guionnet. On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. H. Poincaré Probab. Statist.*, 37(2) :155–194, 2001.

[19] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.

[20] I. H. Dinwoodie and S. L. Zabell. Large deviations for exchangeable random vectors. *Ann. Probab.*, 20(3) :1147–1166, 1992.

[21] C. Dombry. A stochastic model for dna denaturation. *Journal of Statistical Physics*, 120(3-4) :695–719, 2005.

[22] C. Dombry. A weighted random walk model - application to a genetic algorithm. *Preprint*, 2005.

[23] C. Dombry, N. Guillotin-Plantard, B. Pinçon, and R. Schott. Data structures with dynamical random transitions. *Random Structures and Algorithms*, To appear.

[24] R.S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics*. Springer-Verlag, New York, 1985.

[25] W. Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York, 1971.

[26] P. Flajolet, J. Françon, and J. Vuillemin. Sequence of operations analysis for dynamic data structures. *J. Algorithms*, 1(2) :111–141, 1980.

[27] J. Françon, B. Randrianarimanana, and R. Schott. Analysis of dynamic algorithms in D. E. Knuth's model. In *CAAP '88 (Nancy, 1988)*, volume 299 of *Lecture Notes in Comput. Sci.*, pages 72–88. Springer, Berlin, 1988.

[28] A. Greven and F. den Hollander. Large deviations for a random walk in random environment. *Ann. Probab.*, 22(3) :1381–1428, 1994.

[29] N. Guillotin. Asymptotics of a dynamic random walk in a random scenery. II. A functional limit theorem. *Markov Process. Related Fields*, 5(2) :201–218, 1999.

[30] N. Guillotin. Asymptotics of a dynamic random walk in a random scenery I. A law of large numbers. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 36(2) :127–151, 2000.

[31] N. Guillotin-Plantard. Dynamic $\mathbb{Z}^d$-random walks in a random scenery : a strong law of large numbers. *J. Theoret. Probab.*, 14(1) :241–260, 2001.

[32] N. Guillotin-Plantard and R. Schott. Distributed algorithms with dynamical random transitions. *Random Structures Algorithms*, 21(3-4) :371–396, 2002.

[33] J.H. Holland. *Adaptation in natural and artificial systems.* University of Michigan Press, Ann Arbor, Mich., 1975. An introductory analysis with applications to biology, control, and artificial intelligence.

[34] D. Kessler, H. Levine, D. Rigdway, and L. Tsimring. Evolution on a smooth landscape. *J. Statist. Phys*, 1997.

[35] D. E. Knuth. Deletions that preserve randomness. *IEEE Trans. Software Engrg.*, SE-3(5) :351–359, 1977.

[36] L. Kuipers and H. Niederreiter. *Uniform distribution of sequences.* Wiley-Interscience [John Wiley & Sons], New York, 1974. Pure and Applied Mathematics.

[37] J. T. Lewis, C.-E. Pfister, and W. G. Sullivan. Entropy, concentration of probability and conditional limit theorems. *Markov Process. Related Fields*, 1(3) :319–386, 1995.

[38] J. T. Lewis, C.-E. Pfister, and W.G. Sullivan. The equivalence of ensembles for lattice systems : some examples and a counterexample. *J. Statist. Phys.*, 77(1-2) :397–419, 1994.

[39] J.T. Lewis and C.E. Pfister. Thermodynamic probability theory : Some aspects of large deviations. *Russian Math. Surveys*, 50(2) :279–317, 1995.

[40] W. Li. Are isochore sequence homogeneous ? *Gene*, 2002.

[41] G. Louchard. Random walks, Gaussian processes and list structures. *Theoret. Comput. Sci.*, 53(1) :99–124, 1987. Eleventh colloquium on trees in algebra and programming (Nice, 1986).

[42] G. Louchard, C. Kenyon, and R. Schott. Data structures' maxima. *SIAM J. Comput.*, 26(4) :1006–1042, 1997.

[43] G. Louchard, B. Randrianarimanana, and R. Schott. Dynamic algorithms in d.e. knuth's model : a probabilistic analysis. *Theoretical Computer Science*, 93 :201–225, 1991.

[44] R. S. Maier. A path integral approach to data structure evolution. *J. Complexity*, 7(3) :232–260, 1991.

[45] C. Mazza. Strand separation in negatively supercoiled DNA. *Journal of Mathematical Biology*, 51(2) :198–216, 2005.

[46] C. Mazza and D. Piau. On the effect of selection in genetic algorithms. *Random Structures and Algorithms*, 18(2) :185–200, 2001.

[47] J.L. Olivier, P. Carena, M. Heisenberg, and P. Barnaul-Galvan. Islander : computational prediction of isochores in genome sequences. *Nucleid Acids Research*, 2004.

[48] C.F. Osgood, editor. *Diophantine approximation and its applications.* Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1973.

[49] Y. Rabinovich and A. Wigderson. Techniques for bounding the convergence rate of genetic algorithms. *Random Structures and Algorithms*, 14(2) :111–138, 1999.

[50] H. Sun, M. Mezei, R. Fye, and C.J. Benham. Monte carlo analysis of conformational transitions in superhelical dna. *J. Chem.*, 103 :8653–8665, 1995.

[51] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, 81 :73–205, 1995.

[52] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3) :505–563, 1996.

[53] L. Tsimring, H. Levine, and D. Kessler. Rna virus evolution via fitness-space model. *Phy. Rev. Lett.*, 1996.

[54] S. R. S. Varadhan. Asymptotic probabilities and differential equations. *Comm. Pure Appl. Math.*, 19 :261–286, 1966.

[55] S.R.S. Varadhan. *Large deviations and applications.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1984. CBMS-NSF Regional Conference Series in Applied Mathematics.

[56] M. Vose. *The Simple Genetic Algorithm. Foundations and theory.* MIT Press, Cambridge MA, 1999.

[57] C.T. Zhang and R. Zhang. Isochore structures in the mouse genome. *Genomics*, 2004.