UFR S.T.M.I.A.
École Doctorale IAE + M
Université Henri Poincaré - Nancy I
D.F.D. Mathématiques

Thèse

présentée pour l'obtention du titre de

Docteur de l'université Henri Poincaré, Nancy-I

en **Mathématiques**

par **Marie-Pierre ETIENNE**

Sujet: Génération de page de couverture de thèse.

soutenue le plus tôt possible

devant le jury composé de Messieurs les Professeurs:

Président :	Son nom	Du pays d'à coté.
Rapporteurs :	Bernard PRUM	Du pays d'à coté.
	Son nom	Du pays d'à coté.
Examinateurs :	Dominique CELLIER	Université de Rouen
	Jean-Jacques DAUDIN	INAPG
	Son nom	Du pays d'à coté.

Institut Élie Cartan Nancy
INRIA Lorraine

Remerciements

Table des matières

Introduction générale

L'ADN d'un être vivant contient toutes les informations nécessaires à son développement. C'est une molécule complexe mais composée de quatre bases élémentaires. La succession de ces bases forme un code : le code génétique. Le séquençage d'un génome consiste à déterminer cette succession de bases, il reste ensuite à traduire le code, c'est l'enjeu actuel puisque le séquençage du génome humain est presque fini.

Le travail présenté ici est divisé en 3 parties. La première a pour but de définir les notions de bases en biologie nécessaires pour comprendre la problématique à laquelle on s'est intéressé. Cette partie essaie aussi de mettre l'accent sur les difficultés rencontrées lors de l'analyse de séquences biologiques. Elle montre également l'intérêt que présente une étude poussée du score local.

La seconde partie est dédiée à l'étude du score local proprement dit.

Première partie

Biologie et probabilités

Chapitre 1

Quelques éléments de biochimie

Le but de cette première partie est de définir les éléments de biologie sur lesquels nous allons travailler et d'identifier certains des problèmes rencontrés lorsque l'on travaille sur des séquences biologiques. Pour rédiger cette partie je me suis inspirée de [Wat95], [DEKM98] ainsi que de la thèse de Sabine Mercier [Mer99].

PRESENTER LE PLAN

1.1 ADN, ARN et protéines

Une séquence d'ADN, d'ARN ou de protéines est une suite d'éléments fondamentaux. Ces éléments sont au nombre de quatre pour l'ADN ou l'ARN et de vingt pour les protéines.

1.1.1 L'ADN : Acide Désoxyribonucléique

Cette molécule, découverte par Watson et Crick en 1953, contient la totalité de l'information biochimique vitale d'un individu. Elle est formée de deux brins complémentaires qui s'agencent en formant une double hélice, très caractéristique de cette molécule. Un brin est une succession de petites molécules appelées nucléotides. Ils sont constitués d'un groupement phosphate, d'un sucre et d'une base organique, c'est cette dernière qui les différencie les uns des autres.

Lorsqu'on veut décrire une séquence d'ADN, il suffit donc de donner la succession en ces quatre éléments fondamentaux, on les note A (adénine), C (cytosine), G (guanine) et T (thymine).

Les deux brins d'ADN ont des caractéristiques intéressantes : ils sont complémentaires et anti-parallèles. La complémentarité se traduit de la façon suivante : lors de l'appariement des deux brins, on a toujours un A face à T

et un C face à G. De plus les nucléotides se lient les uns aux autres dans un sens précis, et les deux brins d'ADN sont orientés selon des directions opposées : c'est l'anti-parallélisme. On peut résumer ces deux propriétés sur le schéma 1.1 :

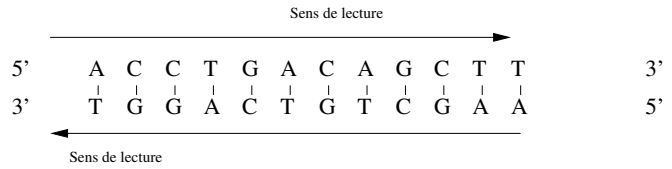

FIG. 1.1 – *Complémentarité et anti-parallélisme de l'ADN.*

1.1.2 l'ARN : Acide Ribonucléique

L'ARN est assez semblable à l'ADN mais il ne comporte qu'un unique brin. Cette molécule a tendance à se replier sur elle-même pour former une structure hélicoïdale. L'ARN se décrit également à partir d'un alphabet à quatre lettres { A, U, G, C }. Au contraire de l'ADN, il y a différents types d'ARN, que l'on distingue par leur fonction ARNm (ARN messager), ARNt (ARN de transfert), ...

1.1.3 Les protéines

Elles sont à la fois des éléments de structure et des éléments fonctionnels de l'organisme d'un être vivant. Elles sont composées d'acides aminés qui sont au nombre de 20.

La synthèse des protéines regroupe une succession de mécanismes biologiques mettant en jeu les différents ARN, elle permet la production de protéines à partir du code génétique contenu dans l'ADN.

Acide Aminé	Abr	Acide Aminé	Abr
alanine	Ala	méthionine	Met
cystéine	Cys	aspargine	Asn
acide aspartique	Asp	proline	Pro
acide glutaminique	Glu	glutamine	Gln
phénylalanine	Phe	arginine	Arg
glycine	Gly	sérine	Ser
histine	His	thréonine	Thr
isoleucine	Ile	valine	Val
lysine	Lys	tryptophane	Trp
leucine	Leu	tyrosine	Tyr

TAB. 1.1 – *Les 20 acides aminés*

Les trois types de macromolécules qui viennent d'être décrits sont à la base de tout organisme vivant. La partie suivante détaille les relations étroites qui les lient et essaie de faire ressortir les principaux problèmes liés à leur étude.

1.2 De l'ADN aux protéines

La synthèse de protéines s'effectue aux travers de 2 étapes principales.

1.2.1 Transcription : de l'ADN aux divers ARN

Les parties codantes de l'ADN sont les portions qui vont donner lieu à la synthèse de protéines. En effet, tout l'ADN n'est pas utile dans ce cadre, on estime à 30% la proportion de régions codantes sur l'ADN humain. Dans un premier temps, l'ADN est transcrit en ARN, par complémentarité et dans un sens bien précis, du site 5' vers le site 3'. Chaque type d'ARN produit a un rôle bien défini, seul l'ARN messager porte le code génétique.

1.2.2 Traduction : de l'ARNm aux protéines

Ce mécanisme met en jeu des ARN différents qui se fixent sur l'ARNm pour traduire chaque codon (groupement de trois nucléotides) en acides aminés selon la correspondance donnée par le tableau 1.2.

première position	**deuxième position**				**troisième position**
	A	**C**	**G**	**T**	
A	AAA Lys	ACA Thr	AGA Arg	ATA Ile	**A**
	AAC Asn	ACC Thr	AGC Ser	ATC Ile	**C**
	AAG Lys	ACG Thr	AGG Arg	ATG Met	**G**
	AAT Asn	ACT Thr	AGT Ser	ATA Ile	**T**
C	CAA Gln	CCA Pro	CGA Arg	CTA Leu	**A**
	CAC His	CCC Pro	CGC Arg	CTC Leu	**C**
	CAG Gln	CCG Pro	CGG Arg	CTG Leu	**G**
	CAT His	CCT Pro	CGT Arg	CTA Leu	**T**
G	GAA Glu	GCA Ala	GGA Gly	GTA Val	**A**
	GAC Asp	GCC Ala	GGC Gly	GTC Val	**C**
	GAG Glu	GCG Ala	GGG Gly	GTG Val	**G**
	GAT Asp	GCT Ala	GGT Gly	GTA Val	**T**

T	TAA Stop	TCA Ser	TGA Stop	TTA Leu	**A**
	TAC Tyr	TCC Ser	TGC Cys	TTC Phe	**C**
	TAG Stop	TCG Ser	TGG Trp	TTG Leu	**G**
	TAT Tyr	TCT Ser	TGT Cys	TTA Phe	**T**

TAB. 1.2 – *Correspondance codons ↔ acides aminés*

1.3 Evolution des séquences

Le matériel génétique évolue au cours du temps de plusieurs manières différentes. Comme nous l'avons déjà signalé, des erreurs de transcription ou de traduction peuvent se produire : généralement elles sont sans conséquences mais elles peuvent parfois empêcher la synthèse d'une protéine ou en produire une quelque peu différente. Un autre type d'évolution est lié à la réplication de l'ADN lui-même : lorsque des erreurs se produisent au cours de sa reproduction et que le code génétique qu'il contient permet encore la synthèse de protéines viables, alors cet ADN est conservé et donne parfois naissance à des espèces mutantes. Les altérations de l'ADN sont multiples, on s'intéresse essentiellement à deux d'entre elles : les substitutions et les indels.

1.3.1 Substitutions

On appelle **substitution** le fait de remplacer un nucléotide par un autre. Par exemple, la suite A ci-dessous a subi une substitution pour donner la suite B.

A : C G **C** T A C T A T G
B : C G **G** T A C T A T G

1.3.2 Délétions et insertions

Un autre type d'altérations du code génétique est la suppression d'un nucléotide. C'est-à-dire qu'on passe de la séquence A à la séquence B en supprimant un nucléotide.

EXEMPLE 1 Voici un exemple de délétion. On obtient la séquence **B** à partir de la séquence **A** en supprimant le troisième nucléotide.

A : C G **C** T A C T A T G
B : C G T A C T A T G

L'opération inverse, consistant à insérer un nucléotide est également possible. En pratique, lorsque l'on observe deux séquences comme A et B ci-dessus, on est évidemment incapable de savoir si une base de A a été supprimée pour donner B ou si on a ajouté une base à B pour obtenir A. On rassemble donc

ces deux types de mutations sous le nom d'insertions-délétions ou encore **indels**.

Ces éléments biologiques sont essentiels à la construction d'un être vivant. Elles régissent non seulement la construction mais également le fonctionnement de tout organisme vivant. Ainsi l'analyse de ces molécules est une étape primordiale dans la connaissance et la compréhension des mécanismes biologiques.

Chapitre 2

Analyse de séquences

Comprendre le rôle de chacune des macromolécules citées dans la partie 1.1 est fondamental pour comprendre les mécanismes du vivant. On aimerait savoir quelle partie d'ADN est codante, à quel gène correspond telle protéine, quelle est la fonction d'une protéine particulière, etc ...

Pour répondre à ces questions, on peut mettre en place des protocoles expérimentaux, mais ceux-ci sont longs et coûteux, d'autant plus qu'on ne sait pas toujours dans quelle direction chercher. L'analyse informatique de séquences a pour but de réduire le domaine de recherche en donnant des pistes de réponses. Par exemple cette protéine dont la fonction est inconnue ressemble à telle autre qui est impliquée dans tel mécanisme.

Analyser des séquences signifie extraire des propriétés de celles-ci, avec pour seule donnée leur succession en éléments fondamentaux. On peut procéder de plusieurs manières différentes : on compare deux séquences, dont l'une est connue, pour savoir si elles se ressemblent et déduire des informations sur la séquence inconnue grâce à la connaissance de la seconde ou on compare tout un groupe de séquences de façon à faire ressortir des motifs communs. On peut aussi travailler sur une seule séquence pour déterminer des propriétés intrinsèques de celle-ci.

2.1 La comparaison de séquences

Lorsque l'on compare deux séquences, on aimerait mettre en évidence des éléments pour conclure à l'existence d'un ancêtre commun. Cet ancêtre aurait évolué, suivant les mécanismes de mutations décrits dans la partie 1.3 pour finalement donner deux séquences différentes mais qui ont gardé des fonctions similaires. Il est donc important de donner un sens à l'expression “ces deux séquences se ressemblent”.

2.1.1 La détermination d'un score

Pour mesurer la similarité entre deux séquences, on calcule un score. Il caractérise soit la similarité soit la dissimilarité entre deux séquences. Ce score repose sur un système qui attribue un score élémentaire pour chaque position lorsque les deux séquences sont écrites l'une sous l'autre comme sur la figure suivante.

A :	G	C	T	G	A	T	T	A	G	C	T
B :	G	G	T	G	A	T	T	A	G	C	T

Le score élémentaire est un élément d'une matrice de score qui présente toutes les possibilités d'appariement. Pour les acides nucléiques, la matrice la plus utilisée est la matrice identité.

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Cette matrice nous indique que l'on attribue un score élémentaire de 1 si les deux bases sont identiques et 0 sinon. Le cas d'un mauvais appariement (score élémentaire de 0) correspond à une substitution à un moment de l'évolution.

On se rend compte rapidement qu'un décalage peut permettre de mieux mettre en évidence des similarités entre séquences. Une insertion ou une délétion d'une ou plusieurs bases permet de mettre en évidence des zones similaires. Ces brèches doivent etre prises en compte dans le score et pénalisées si l'on cherche à quantifier la similarité des séquences. Elles correspondent à des indels durant une phase d'évolution.

Si l'on cherche à déterminer une similarité entre deux séquences, on calcule un score de la manière suivante :

$$S = \sum s_{ele} - \sum s_b,$$

où s_{ele} désigne le score élémentaire d'un appariement et s_b le score d'une brèche.

Nous reviendrons par la suite sur la justification d'un système de score additif.

Dans cette section nous nous sommes intéressés au score de deux séquences alignées. Mais l'alignement de deux séquences pose de vraies questions qui sont abordées dans le paragraphe suivant.Comment choisir un bon alignement ?

2.1.2 Alignement de deux séquences

Pour déterminer la similarité de deux séquences, nous avons jusque là uniquement évoqué la méthode du score; en fait, il existe plusieurs critères de similarité entre 2 séquences; on peut par exemple s'intéresser au plus long segment commun : il donne la plus grande portion conservée entre les deux séquences. Un autre outil est la distance de Levenshtein qui compte le nombre d'opérations élémentaires (substitutions ou délétions) nécessaires pour passer d'une séquence à l'autre. Mais nous nous contenterons ici d'utiliser le score pour estimer la similarité de deux séquences, c'est en fait la méthode la plus utilisée.

Aligner deux séquences consiste à les écrire l'une au-dessus de l'autre de façon à mettre en évidence leurs similarités, toujours selon les 3 opérations élémentaires.

EXEMPLE 2 Voici un exemple d'alignement tiré de [Wat95]. Sur la troisième ligne, on a représenté les nucléotides communs. On cherche un alignement possible pour les séquences **A** =GCTGATATAGCT et **B** = GGGTGATTAGCT.

A :	-	G	C	T	G	A	T	A	T	A	G	C	T
B :	G	G	G	T	G	A	T	-	T	A	G	C	T
			G	T	G	A	T		T	A	G	C	T

Dans cet exemple, on a effectué un alignement global, c'est-à-dire qu'on a aligné la totalité des deux séquences. Il est également possible de considérer des alignements locaux, on regarde alors des alignements de parties contigües de chacune des deux séquences. Nous y reviendrons par la suite.

Alignement global

Pour obtenir un alignement global, on procède de la manière suivante : étant données deux séquences A=$a_1\, a_2\, \ldots\, a_n$ et B=$b_1\, b_2\, \ldots\, b_m$, on commence par insérer des "trous" de façon à ce qu'elles aient la même longueur L. On écrit ensuite la suite A*=$a_1^*\, a_2^*\, \ldots\, a_L^*$ obtenue au-dessus de la suite B*=$b_1^*\, b_2^*\, \ldots\, b_L^*$ (un élément étoilé est soit un élément de la suite de départ, soit un trou (une brèche) symbolisé par (-). On obtient ainsi l'alignement suivant :

A* :	a_1^*	a_2^*	$\ldots$	a_L^*
B* :	b_1^*	b_2^*	$\ldots$	b_L^*

Le nombre d'alignements possibles est $\mathcal{C}_{n+m}^m$ où n et m désignent les longueurs respectives de A et B. Parmi ce grand nombre d'alignements, on aimerait déterminer les bons alignements. Il n'y a pas de définition intrinsèque d'un bon alignement, cette notion dépend complètement des critères choisis. On peut, par exemple chercher à les aligner de façon à faire apparaître le plus

long segment commun évoqué ci-dessus. Une autre solution, très utilisée est d'utiliser un système de score comme défini précédemment. A chaque paire d'éléments $(\mathrm{a}_i^*, \mathrm{b}_i^*)$, on associe un score $s(\mathrm{a}_i^*, \mathrm{b}_i^*)$: le score élémentaire; le score d'alignement noté $S(\mathbf{A}^*, \mathbf{B}^*)$ est la somme des scores de chaque paire.

Exemple 3 On donne ici un exemple de modèles de score. On définit la fonction de score s à valeurs dans $\mathbb{R}$, telle que :

$$s(a_i, b_j) = \begin{cases} 1 & \text{si } a_i = b_j \\ 0 & \text{sinon} \end{cases} .$$

Ce choix d'une fonction de score correspond au choix de la matrice unité pour la détermination des scores élémentaires et à une absencede pénalisation pour les brèces.

Dans ce cas très simple, le score d'un alignement sera le nombre de bases communes à chacune des deux séquences et à la même position. On reprend l'alignement donnée dans l'exemple 2.

A :	-	G	C	T	G	A	T	A	T	A	G	C	T
B :	G	G	G	T	G	A	T	-	T	A	G	C	T
	0	1	0	1	1	1	1	0	1	1	1	1	1

Le score de cet alignement est donc 10.

Le choix d'une fonction de score, et par conséquent de la matrice de score associée, est déterminant. C'est un problème à part entière. En effet chercher un alignement optimal consiste à maximiser une fonction de score, il faut donc adapter celle-ci aux caractéristiques que l'on veut mettre en évidence.

L' alignement otimal des séquences $\mathbf{A}$ et $\mathbf{B}$ sous la fonction de score s est donc :

$$\mathrm{argmax}_{\mathbf{A}^*, \mathbf{B}^*} S(\mathbf{A}^*, \mathbf{B}^*).$$

La valeur du score d'alignement optimal est le **score global** de la séquence noté $S(\mathbf{A}, \mathbf{B})$.

Etant donné le grand nombre d'alignements possibles, déterminer le meilleur alignement nécessite une réflexion algorithmique pour obtenir des temps de calcul raisonnables. Il existe différents algorithmes de recherche d'alignement optimal. Le plus connu est l'algorithme de Needleman-Wunsch qui utilise la programmation dynamique, il ne prend en compte que des fonctions de score pour lesquelles la pénalisation d'un indel est proportionnelle à sa longueur, des raffinements existants autorisent la pénalisation d'indels par une fonction affine. On trouve aussi des algorithmes heuristiques, qui obtiennent de bons alignements mais qui n'assurent pas leur optimalité. C'est le cas du programme BLAST qui commence par chercher des ancres d'alignements (des points de repère avec un fort taux de similarité) et qui étend ensuite l'alignement à partir de ces ancres.

Alignement local

On a vu que l'on peut aussi s'intéresser à un alignement local, c'est-à-dire que l'on considère un alignement de sous-séquences. Le principe de base reste le même.

EXEMPLE 4 On reprend les séquences données dans l'exemple 2, un alignement local pourrait être le suivant :

A :	T	A	G	C	T
B :	T	A	G	C	T

Le but de ces alignements locaux est de détecter des régions similaires sans connaissances a priori des longueurs de zones à considérer. L'alignement local comporte donc une partie de chacune des séquences et non pas les séquences dans leur globalité.

Bien que l'idée de départ soit la même, le problème informatique est nettement plus compliqué puisque l'on considère tous les alignements de toutes les sous-séquences possibles, les possibilités sont donc bien plus nombreuses encore.

Comme pour un alignement global, on va chercher à déterminer l'alignement local optimal, c'est-à-dire celui qui maximise la fonction de score choisie. Lorsque les fonctions de score choisies sont suffisamment simples (linéaires ou affines), il existe des algorithmes efficaces de recherche d'alignement local optimal, par exemple celui de Smith et Watermann. Ces algorithmes reposent le plus souvent sur des principes de programmation dynamique.

2.1.3 Alignement multiple

Il est parfois important notamment dans le cas de séquences d'acides aminés de considérer l'alignement simultané de plusieurs séquences et non pas seulement l'alignement deux à deux. Nous reviendrons plus longuement sur les difficultés qui découlent de ce type d'alignement dans le premier chapitre de la troisième partie.

2.2 L'étude d'une unique séquence

Il est souvent utile d'étudier une séquence pour elle-même dans le but de rechercher des caractéristiques intrinsèques telles que la charge, le volume. On peut définir les notions de score local ou global évoquées ci-dessus pour une unique séquence. A nouveau le choix de la fonction de score est prépondérant puisque l'on va s'intéresser encore une fois au segment réalisant le score maximal.

Evidemment la fonction de score utilisée n'est plus la même. C'est une fonction qui va de l'alphabet $\mathcal{A}$ dans $\mathbb{R}$ (et non plus de $\mathcal{A}^2$ dans $\mathbb{R}$). On assigne un score élémentaire à chaque élément de la suite considérée en fonction des propriétés auxquelles on s'intéresse. Le score global de la séquence est la somme des scores élémentaires et le score local est le maximum des scores de toutes les sous séquences (c.f. la section 3.2 pour une définition plus précise du score local).

Notons tout de même que le problème informatique de l'étude d'une seule séquence diffère un peu de celui de l'étude d'un alignement. En effet, on ne définit qu'un seul score global pour la séquence : c'est le score de la séquence. Si l'on cherche à identifier une partie singulière de cette séquence, on va alors utiliser le score local avec une fonction de score judicieusement choisie, de façon à ce que la portion de séquence qui réalisera le meilleur score local soit la portion singulière qui nous intéresse. Par exemple lorsque l'on cherche à identifier une partie hydrophobe d'une protéine, on choisira une fonction de score qui pénalise les acides aminés hydrophiles et au contraire récompense les autres. Le sous-segment réalisant le meilleur score désignera la partie la plus hydrophobe de la protéine considérée.

2.3 Retour sur le score

Cette section justifie le choix d'un modèle de score comme décrit ci-dessus et donne quelques éléments sur la construction de matrices de score.

2.3.1 Une justification du système de score additif

On part de l'hypothèse que l'évolution est markovienne; ce qui signifie que l'évolution d'une séquence biologique à une date donnée ne dépend pas de toute son histoire mais uniquement de son état à l'instant considéré. On suppose également que les sites de mutation sont indépendants. On considère deux séquences biologiques x et y de longueurs respectives n et m. On appelle x_i (respectivement y_j) le $i^{\text{ème}}$ (resp. $j^{\text{ème}}$) symbole de la séquence x (resp. y). Ces symboles x_i ou y_j sont issus d'un alphabet $\mathcal{A}$ qui serait { A, C, G, T } dans le cas de séquences nucléotidiques et l'ensemble des 20 acides aminés dans le cas de l'étude des protéines. On va considérer maintenant un alignement global sans brèche et deux séquences de même taille.

On cherche à savoir si les séquences dérivent l'une de l'autre ou si elles n'ont aucun lien entre elles. On traduit cette alternative de la manière suivante :

- Si les deux séquences n'ont pas d'ancêtres communs, alors la probabilité d'observer x_i en face de y_i est donnée par $q_{x_i} \times q_{y_j}$. C'est ce qu'on appellera le modèle aléatoire (R). Ainsi la probabilité d'observer l'ali-

gnement est donnée par

$$\mathbb{P}(x, y/R) = \prod_i q_{x_i} \prod_j q_{y_j}. \tag{2.3.1}$$

– Si les deux séquences dérivent l'une de l'autre, on se place dans le cadre de l'évolution markovienne. Alors la probabilité que y_i dérive de x_i sous le modèle markovien (M) est $q_{x_i}\pi_{x_i y_i}$ où q_{x_i} et la probabilité d'observer x_i est $\pi_{x_i y_i}$ est la probabilité de passer de x_i à y_i dans le modèle (M). Ainsi la probabilité d'observer l'alignement est donnée par

$$\mathbb{P}(x, y/M) = \prod_i q_{x_i}\pi_{x_i y_i} = \prod_i p_{x_i y_i}. \tag{2.3.2}$$

Comme il est impossible de savoir si x dérive de y ou y dérive de x on choisit q et π de façon à ce que pour toute lettre a, b de l'alphabet $\mathcal{A}$ on ait la relation suivante :

$$q_a\pi_{ab} = q_b\pi_{ba} = p_{ab} = p_{ba} \qquad \forall (a, b) \in \mathcal{A}^2. \tag{2.3.3}$$

Afin de décider si l'alignement des deux séquences considérées met en évidence des ressemblances, on forme le quotient de ces deux quantités, c'est-à-dire :

$$\frac{\mathbb{P}(x, y/M)}{\mathbb{P}(x, y/R)} = \frac{\prod_i q_{x_i}\pi_{x_i y_i}}{\prod_i q_{x_i} \prod_j q_{y_j}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}. \tag{2.3.4}$$

Pour obtenir un système additif pour le score, il suffit de passer au logarithme. Le score global de l'alignement sera alors donné par :

$$S(x, y) = \sum_i s(x_i, y_i), \tag{2.3.5}$$

où

$$s(a, b) = \ln \frac{p_{ab}}{q_a q_b}.$$

Nous allons voir dans la section suivante que les matrices de score (ou encore matrices de substitution) sont déterminées en calculant des logarithmes de fréquences relatives, ce qui justifie bien le modèle de score adopté.

Bien sûr, dans cette analyse nous avons supposé que tous les sites de mutations sont indépendants et que l'alignement était donné. Mais cette vision simplifiée permet de bien comprendre la construction des matrices de score détaillée ci-dessous.

2.3.2 Les matrices de score

On admet généralement qu'un acide aminé peut être substitué à un autre ayant des propriétés physico-chimiques similaires sans que que la structure ou la fonction d'une protéine en soit modifiée. Il faut donc en tenir compte lorsque l'on considère une substitution : **toutes les substitutions n'ont pas le même impact sur l'évolution d'une protéine.** On regoupe donc les acides aminés en famille de manière à définir un système de score qui tient compte de ces similitudes.

Matrices PAM

Elles font partie des matrices les plus utilisées. Elles représentent les mutations possibles d'un acide aminé lors de l'évolution des protéines très semblables (moins de 15% de différence) que l'on pouvait facilement aligner. L'alignement de ces protéines a été effectué en utilisant le principe de parcimonie : on a construit un arbre dans lequel on cherche à expliquer l'évolution avec un minimum de substitutions. A partir de ces alignements, on a calculé une matrice de probabilité dans laquelle chaque élément donne la probabilité qu'un acide aminé A soit remplacé par un acide aminé B durant une étape d'évolution. Cette matrice de substitution correspond au fait que l'on cherche les protéines pour lesquelles l'activité n'a pas été détruite ce qui correspond à admettre en moyenne une substitution pour 100 sites durant un temps particulier d'évolution. On appelle cette matrice une 1PAM (Percent Accepted Mutations) matrice. Regarder cette matrice élevée à la puissance x (matrice xPAM) consiste à regarder les probabilités de mutation sur une durée x. Les matrices utilisées sont en fait les matrices XPAM (matrices de mutation de Dayhoff): elles sont obtenues en calculant le logarithme des fréquences relatives de mutations de chaque acidé aminé. La 250PAM semble être la matrice la mieux adaptée pour distinguer les protéines apparentées (étude de [Sch79]).

Cette matrice, bien que très utilisée, présente quelques inconvénients. Elle considère que les mutations ne dépendent pas du site où elles se produisent et les protéines prises en compte pour l'étude ne sont pas représentatives de toutes les classes de protéines connues aujourd'hui.

Matrices BLOSUM

Ces matrices sont apparues après les matrices de Dayhoff pour faire apparaître des similarités entre des séquences plus éloignées. Elles ont été construites à partir d'un ensemble de régions protéiques alignées sans brèches. Les protéines obtenues sont alors regroupées dans des classes; des protéines de la même classe possède au moins L% d'identités. On compte alors le nombre N_{ab} de résidus a d'un cluster aligné avec un résidu b dans un autre cluster. Ce nombre est rectifié pour tenir compte de la taille des clusters et

chaque N_{ab} est divisé par $n_1 \times n_2$ où n_1 et n_2 désignent la taille des clusters considérés. On détermine ensuite la probabilité d'observer un a par

$$q_a = \frac{\sum_b N_{ab}}{\sum_{c,d} N_{cd}},$$

et la probabilité d'observer un a aligné avec un b par

$$p_{ab} = \frac{N_{ab}}{\sum_{c,d} N_{cd}}.$$

Nous avons vu que pour améliorer un alignement de séquences, on pouvait insérer des brèches. Si on souhaite que le score le prenne en compte, il est nécessaire de pénaliser cette insertion.

2.3.3 Les pénalités pour les brèches

Il existe essentiellement deux systèmes de pénalisation.
1) La pénalité est proportionnelle à la longueur de la brèche.

$$P = kL,$$

où L est la longueur de la brèche.
2) la pénalité est une fonction affine de la longueur.

$$P = kL + m.$$

Souvent $m = 10k$, ce qui signifie que c'est l'introduction d'une brèche qui est fortement pénalisante, la longueur de celle-ci prend alors moins d'importance. Ce type de pondération des brèches est relativement bien justifié biologiquement; en effet on observe beaucoup plus souvent de longues insertions ou délétions plutôt que de nombreuses petites.

On pourrait bien sur imaginer d'autres systèmes de pénalisation; néanmoins les deux systèmes prśentés sont ceux qui sont utilisés en pratique dans les algorithmes d'alignement.

Chapitre 3

Significativité statistique

Comparer des séquences biologiques se ramène en fait à la comparaison de chaînes de caractères. Il est néanmoins important de pouvoir décider si ce que l'on met en évidence est susceptible d'avoir une signification biologique ou si seul le hasard peut en être responsable. Pour cela il est important d'étudier la significativité statistique des résultats obtenus.

Le problème de la significativité statistique a été bien étudié dans le cas du score global. Le première partie présente les méthodes empiriques, qui sont des méthodes théoriquement simples et faciles à mettre en oeuvre. Nous présenterons également deux autres approches plus théoriques : l'approche bayésienne et l'approche par les valeurs extrêmes. Enfin la dernière partie s'attachera spéfiquement au score local en essayant de bien définir le modèle probabiliste et surtout les outils utilisés.

3.1 Approche empirique

3.1.1 Le pourcentage d'identités

La plus simple consiste à mesurer le pourcentage d'identités entre les deux séquences. C'est l'idée naturelle : si des séquences sont proches, elles ont beaucoup de bases communes et ceci doit pouvoir se mesurer en terme de pourcentage d'identité. Il faut néanmoins être vigilant avec ce critère. Le seuil à partir duquel des séquences peuvent être considérées comme similaires dépend de la nature des séquences considérées. Ainsi des séquences protéiques de 100 résidus ayant au moins 25% d'identités ont vraissemblablement un ancêtre commun tandis que deux séquences nucléiques de 100 bases ayant plus de 50% d'identités n'ont pas forcément de lien biologique. cela provient du fait qu'un base nucléotidique a une fréquence d'apparition bien plus élevée.

3.1.2 Le Z-score

Cette méthode est assez simple et rapide à mettre en oeuvre, elle intègre la notion de score. L'idée est la suivante : on prend l'une des deux séquences et on engendre des séquences aléatoires à partir de celle-ci en mélangeant les lettres de la séquence (technique du bootstrap). Cette méthode permet de garder la même composition en bases que la séquence initiale. On aligne alors les séquences aléatoires ainsi obtenues et on calcule le score associé. On obtient ainsi une distribution empirique pour le score. On compare ensuite le score initial des deux séquences à la distribution du score obtenue grâce aux séquences aléatoires.

En application directe, on définit un deuxième score qui a pour vocation de mesurer l'écartement du score initial par rapport à la distribution aléatoire : le Z-score ([DSO78]).

$$Z = \frac{S - m}{\sigma},$$

où S est le score initial calculé, m la moyenne de la distribution empirique et σ l'écart type de cette même distribution. On évalue donc en fait de combien d'écart types le score obtenu s'éloigne de la moyenne.

Si on suppose que Z suit une loi normale centrée réduite, alors on peut estimer si le Z-score obtenu en pratique est significatif ou non. Mais on sait que cette supposition est rarement justifiée ([Wat95], [KA90]). Le problème de la significativité de ce score n'est donc pas résolu de manière satisfaisante.

3.2 Modélisation

Pour avoir une approche plus rigoureuse de la significativité statistique des résultats obtenus, il est nécessaire de définir un modèle probabiliste pour les séquences. On cherchera ensuite à tester l'hypothèse nulle H_0 qui correspond à l'indépendance des séquences. Il est pour cela essentiel de connaître au mieux la loi du score dans ce cadre.

On va tout d'abord définir le modèle probabiliste étudié. On considère que les sites de mutations sont indépendants, on va donc les modéliser par des variables aléatoires indépendantes à valeurs dans un alphabet fini $\mathcal{A}$.

On a donc deux séquences aléatoires (X) et (Y) de taille respectives n et m, on insère des brèches pour leur donner la même longueur, on obtient deux suites (X^*) et (Y^*), pour lesquelles un élément X_i^* (resp Y_i^*) est soit un élément de la suite d'origine soit une brèche de taille 1. Le dessin 3.1 présente un alignement possible.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	$\dots$	X_{n-1}	X_n	-	-
-	-	-	Y_1	Y_2	-	Y_3	$\dots$	$\dots$	$\dots$	Y_{m-1}	Y_m

TAB. 3.1 – *Un alignement des séquences* (X) *et* (Y)

On se place dans le cas où les brèches sont pénalisées de manière linéaire : la pénalité attribuée à une brèche est proportionnelle à sa longueur. On peut alors définir le score d'un alignement de la manière suivante

$$S = \sum s(X_i^*, Y_i^*), \tag{3.2.1}$$

où s est la fonction de score choisie. Pour un alignement donné, le problème se ramène à l'étude d'une somme de variables aléatoires indépendantes et identiquement distribuées.

Le score global est le plus grand score d'alignement. Il s'écrit sous forme mathématique de la manière suivante :

$$S(X,Y) = \max_{X^*,Y^*} \left(\sum s(X_i^*, Y_i^*) \right). \tag{3.2.2}$$

Cette notation est la traduction mathématique de celle que nous avions adoptée pour définir un alignement, néanmoins pour les calculs; il est préférable d'adopter une notation équivalente. On se donne deux sous-suites strictement croissante u et v de $\{1,\dots,n\}$ et $\{1,\dots,m\}$. Alors le score global de (X) et (Y) est donné par la formule :

$$S(X,Y) = \max_{\substack{1 \leq l \leq \inf(n,m) \\ 1 \leq u(1) < u(2) < \dots < u(l) \leq n \\ 1 \leq v(1) < v(2) < \dots < v(l) \leq m}} \left(\sum s(X_{u(i)}, Y_{v(i)}) - k(n+m-2l) \right), \tag{3.2.3}$$

où k est la pénalisation d'une brèche.

Remarque : Il apparaît clairement sur cette formule que le problème est plus compliqué quand la pénalisation des brèches est une fonction affine, puisqu'ils faut exactement connaître la taille de chaque brèche. Il ne suffit plus de retrancher le nombre total d'indels.

Le score local de deux séquences est alors donné par une formule similaire :

$$H = \max_{\substack{0 \leq i \leq n, 0 \leq j \leq m \\ 1 \leq l \leq n-i \\ 1 \leq p \leq m-j}} \{ S(X_{i+1} \dots X_{i+l}, Y_{j+1} \dots Y_{j+p}) \}. \tag{3.2.4}$$

Il est possible également de définir le score d'une unique séquence (X) et son score local; c'est très utile lorsque l'on s'intéresse à la détection de propriétés intrinsèques de la séquence.

$$S_u(X) = \sum_{1 \leq i \leq n} s_u(X_i), \tag{3.2.5}$$

où s_u est une fonction de score définie sur une unique séquence (cf partie 2.2). Le score local de cette séquence est alors le score maximal de toutes les sous séquences, il se traduit mathématiquement de la manière suivante :

$$H = \max_{\substack{0 \leq i \leq n \\ 1 \leq l \leq n-i}} \{ S_u(X_{i+1} \dots X_{i+l}) \}, \tag{3.2.6}$$

Dans toute la suite et en l'absence de précision, lorsque nous parlerons de score local, il s'agira du score local d'une seule séquence.

Nous allons maintenant détailler deux approches bien distinctes pour l'étude du score global et nous verrons ensuite ce qui est appliqué au score local.

3.3 Approche bayésienne

On cherche à évaluer la probabilité que les deux séquences alignées proviennent d'un ancêtre commun, on va développer ici le point de vue bayésien. Mathématiquement on cherche à évaluer la quantité $\mathbb{P}(M/X,Y)$ c'est-à-dire la probabilité étant données les deux séquences, qu'elles suivent le modèle d'évolution markovien. Rappelons que l'alternative au modèle markovien (qui correspond à l'existence d'un ancêtre commun) est le modèle aléatoire (pas de points communs entre les séquences). On note $\mathbb{P}(M)$ la probabilité que le modèle soit markovien et $\mathbb{P}(R) = 1 - \mathbb{P}(M)$ son alternative.

$$\begin{aligned}
\mathbb{P}(M/X,Y) &= \frac{\mathbb{P}(X,Y/M)\,\mathbb{P}(M)}{\mathbb{P}(X,Y)}, \\
&= \frac{\mathbb{P}(X,Y/M)\,\mathbb{P}(M)}{\mathbb{P}(X,Y/M)\,\mathbb{P}(M) + \mathbb{P}(X,Y/R)\,\mathbb{P}(R)}, \\
&= \frac{(\mathbb{P}(X,Y/M)\,\mathbb{P}(M))\,/\,(\mathbb{P}(X,Y/R)\,\mathbb{P}(R))}{(\mathbb{P}(X,Y/M)\,\mathbb{P}(M))\,/\,(\mathbb{P}(X,Y/R)\,\mathbb{P}(R)) + 1}.
\end{aligned}$$

On pose

$$S' = S + \ln \frac{\mathbb{P}(M)}{\mathbb{P}(R)},$$

où

$$S = \ln \frac{\mathbb{P}(X,Y/M)}{\mathbb{P}(X,Y/R)}.$$

Alors la probabilité qu'on soit dans le cadre d'une évolution de deux séquences à partir d'un ancêtre commun est donnée par :

$$\mathbb{P}(M/X,Y) = \sigma(S'),$$

avec

$$\sigma(x) = \frac{e^x}{1+e^x}.$$

Etant donné un système de score, calculer $\mathbb{P}(X,Y/M)$ ou $\mathbb{P}(X,Y/R)$ ne présente pas de difficultés (c.f. partie 2.3.1).

La difficulté apparaît dans le choix de la valeur a priori de $\ln \mathbb{P}(M)/\mathbb{P}(R)$. Fixons cette valeur. Si on compare une séquence à l'ensemble des séquences d'une base de données, on va en trouver un certain nombre qui semblent

correspondre à notre séquence de départ mais qui ne sont proches que par chance. Par exemple si on décide de prendre 1/9 comme rapport pour $\mathbb{P}(M)/\mathbb{P}(R)$, cela signifie qu'en moyenne lorsque l'on compare notre séquence à dix autres, on va en trouver neuf qui ne présenteront aucun lien et une dont on supposera qu'elle a un ancêtre commun avec la nôtre.

Si on regarde maintenant une banque contenant 100 séquences, il est assez probable d'en trouver 10 ayant un ancêtre commun avec celle qui nous intéresse.

Ainsi la taille de la base de données à laquelle on compare notre séquence joue un rôle essentiel. Le nombre de séquences corrélées avec celle qui nous intéresse augmente linéairement (en moyenne) avec la taille de la base de données.

3.4 Approche par les valeurs extrêmes

Dans cette approche on regarde le maximum de N scores calculés sur des séquences indépendantes. Si la probabilité que ce maximum dépasse le score observé est petite, alors l'observation est considérée comme statistiquement significative.

L'avantage de cette méthode est de prendre en compte la taille de la base de données à laquelle on compare notre séquence.

Dans le cas d'un alignement global sans brèche, le score est la somme de scores élementaires, donc dans notre modèle c'est la somme d'une suite de variables aléatoires indépendantes. Cette somme renormalisée peut être approchée par une variable aléatoire suivant une loi normale.

On compare notre séquence de référence à un ensemble de N séquences d'une base de données et on garde le meilleur score d'alignement. Cela revient mathématiquement à étudier le maximum M_N de N variables aléatoires gaussiennes, la loi de ce maximum est connue; un équivalent pour N grand est donné par :

$$\mathbb{P}\left(M_N \geq x\right) \underset{N\to\infty}{\sim} \exp - KNe^{\lambda(x-\mu)}. \tag{3.4.1}$$

où K, λ et μ sont des constantes qui dépendent de la loi des variables aléatoires initiales.

Ainsi si la probabilité que M_N dépasse le score observé est très faible (de l'ordre de 0.05 ou 0.01), on peut conclure que ce score est statistiquement significatif.

3.5 Les résultats existants pour le score local

Le score local est un outil important puisqu'il permet d'analyser la significativité statistique des alignements locaux obtenus à l'aide des logiciels de comparaison de séquences. Nous énoncerons ici deux résultats connus sur

le comportement du score local. Le résultat de Karlin *et al.* [DK92], [KA90] qui donne une approximation asymptotique du score local et le résultat de Daudin et Mercier [DM99],[Mer99] qui donne la distribution exacte du score locale.

3.5.1 Comportement asymptotique du score local

Nous énonçons ici le résultat de Karlin *et al* [DK92].

Theorem 5 (Approximation de Karlin *et al.*[DK92]) *Soit $(X_i)_{i\geq 1}$ une suite de variables indépendates et identiquement distribuées à valeurs dans $\mathbb{Z}$ d'espérance négative. Soit H_n son score local. Alors*

$$\lim_{n\to+\infty} \mathbb{P}\left(H_n \leq \frac{\ln n}{\lambda} + x\right) = \exp\left(-K^* e^{-\lambda x}\right), \tag{3.5.1}$$

Où λ est l'unique solution positive de l'équation $\mathbb{E}\left[e^{x X_i}\right] = 1$ et K^* une constante qui ne dépend que de la loi de la suite $(X_i)_i$.

Ce résultat est valable lorsque les variables aléatoires considérées pour modéliser le score des bases sont de moyennes négatives. Dans le premier chapitre de la partie 2, nous montrerons que le score local dans le cas d'une moyenne nulle ou proche de 0 croît en $\sqrt{n}$ où n est la longueur des séquences considérées.

En fait nous allons considérer la suite de processus $\left(H^{(N)}\right)_{N\geq 1}$ linéaires par morceaux définis de la façon suivante :

$$\begin{cases} t \mapsto H^{(N)}(t) \text{ est linéaire sur chaque intervalle de la forme } \left[\frac{j}{N}; \frac{j+1}{N}\right] \\ H^{(N)}\left(\frac{j}{N}\right) = \frac{1}{\sqrt{N}} H_j. \end{cases}$$

Dans ce cas, la suite $(X_n)_{n\geq 1}$ sera soit une suite de variables indépendantes centrées ayant un moment d'ordre 2, soit une chaîne de Markov irréductible et stationnaire sur un ensemble fini de $\mathbb{R}$, telle que $\mathbb{E}_\nu(X_1) = 0$.

Dans le premier cas on pose $\sigma^2 = Var(X_1)$, dans le second on suppose que

$$\sigma^2 = \mathbb{E}_\nu(X_1^2) + 2\sum_{k=2}^{\infty} \mathbb{E}_\nu(X_1 X_k), \tag{3.5.2}$$

où ν est la distribution invariante de $(X_n)_{n\geq 0}$.
σ^2 est bien défini car la série (1.2.4) est convergente ([Bil68, p. 166]).

On a alors le résultat suivant pour le score local :

Theorem 6 *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires comme décrit ci-dessus.*
Alors la suite des processus $\left(H^{(N)}(t),\ t \geq 0\right)$ converge en loi vers le processus $(\sigma \max_{0\leq u\leq s} |B_u|, s \geq 0)$, quand N tend vers l'infini.

3.5.2 Distribution du score local

3.6 Conclusion

L'importance de l'étude de la significativité statistique des résultats obtenus est évidente, la formulation des problèmes est assez simple, puisque la question se résume à : "Ce que j'observe peut-il être uniquement dû au hasard ?". Néanmoins la significativité statistique est un problème difficile. Nous nous sommes concentrés sur l'étude d'une unique suite qui semble être une étape indispensable.

Bibliographie

[Bil68] P. Billingsley. *Convergence of probability measures.* John Wiley and Sons, 1968. New York.

[DEKM98] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis.* Cambridge University Press, 1998.

[DK92] A. Dembo and S. Karlin. Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Prob*, 24:113–140, 1992. USA.

[DM99] J.J. Daudin and S. Mercier. Distribution exacte du score local d'une suite de variables indépendantes et identiquement distribuées. *C. R. Acad. Sciences*, 9:815–820, 1999. Série I, Math.

[DSO78] M. O Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model for evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5:345–352, 1978.

[KA90] S. Karlin and Altschul. Methods for assessing the statistical signifiance of molecular sequence features by using general scoring schemes. *Proc. Natl . Acad. Sci*, 87:2264–2268, 1990. USA.

[Mer99] Sabine Mercier. *Statistiques des scores pour l'analyse et la comparaison de séquences biologiques.* PhD thesis, Université de Rouen, décembre 1999.

[Sch79] Schwarz and Dayhoff. Matrices for detecting distant relationships. *Atlas of protein sequences*, pages 353–358, 1979.

[Wat95] Michael S. Waterman. *Introduction to computational biology.* Chapman & Hall, 1995.

Deuxième partie

Le score local

Chapitre 1

Asymptotic behaviour of the local score of independent and identically distributed random sequence

Asymptotic behaviour of the local score of independent and identically distributed random sequences.

Jean-Jacques DAUDIN[a], Marie Pierre ETIENNE[b], Pierre VALLOIS[b].

[a]Institut National Agronomique Paris-Grignon,
Département OMIP, UMR INAPG-INRA, 96021111,
16, rue C. Bernard, 75231 Paris Cedex 05, France.
E-mail : daudin@inapg.inra.fr

[b]Institut de Mathématiques Elie Cartan, Université Henri Poincaré.
BP. 239, 54506 Vandoeuvre Lès Nancy Cedex, France.
E-mail : Marie-Pierre.Etienne@iecn.u-nancy.fr
E-mail : Pierre.Vallois@iecn.u-nancy.fr

17th September 2002

Abstract

Let $(X_n)_{n\geq 1}$ be a sequence of real r.v.'s, we define the local score as $H_n = \max_{1\leq i<j\leq n}(X_i + \ldots + X_j)$. $(X_n)_{n\geq 1}$ is either (a) a sequence of i.i.d. random variables or (b) a "good" Markov chain under its invariant measure. We prove that, if the X_i are centered, $H_n/\sqrt{n}$ converges in distribution to B_1^*, $n \to +\infty$, where $B_1^* = \max_{0\leq u\leq 1}|B_u|$ and $(B_u, u \geq 0)$ is a standard Brownian motion, $B_0 = 0$. In the case (a) if $\mathbb{E}(X_1) = \delta/\sqrt{n}$ and $Var(X_1) = \sigma^2 > 0$, we prove the convergence of $H_n/\sqrt{n}$ to $\sigma\xi_{\delta/\sigma}$ where $\xi_\gamma = \max_{0\leq u\leq 1}\{(B(u)+\gamma u) - \min_{0\leq s\leq u}(B(s)+\gamma s)\}$. We approximate the distribution function of ξ_γ and we determine the asymptotic behaviour of $P(\xi_\gamma \geq a)$, $a \to +\infty$.

Keywords : Brownian motion with drift, local score.
AMS 1991 Subject classifications
60G17, 60G35, 60J15, 60J20, 60J55, 60J65.

1.1 Introduction

1.1 Let $(X_n)_{n\geq 1}$ be a sequence of real valued random variables. We consider $S_n = \sum_{k=1}^n X_k$, $S_0 = 0$, the associated random walk. Let

$H_n = \max_{0\leq i<j\leq n}(S_j - S_i)$ be the local score assigned to $(X_n)_{n\geq 1}$. The aim of this paper is to study the asymptotic behaviour of H_n when $n \to \infty$, $(X_n)_{n\geq 1}$ being either a sequence of i.i.d. random variables or a Markov chain.
The motivations come from biology. The local score is an important tool for DNA sequences analysis. Since the length of DNA is large, the knowledge of the limit behaviour of H_n is actually useful.
Some authors have already studied the local score. In a context of queue theory, Iglehart ([Igl72]) has investigated the convergence of random variables (i.e. virtual waiting time) which looks like the local score.
When $(X_n)_{n\geq 1}$ is either a sequence of i.i.d. rv's or a Markov chain, Daudin and Mercier [DM99] have given an algorithm to determine explicitly $\mathbb{P}(H_n < x)$, for any $x > 0$ and $n \geq 1$. They have introduced a $x \times x-$matrix Π, such that $\mathbb{P}(H_n < x)$ can be expressed via P_n, where P_n is the $x-$dimensional vector : $P_n = P_0\Pi^n$, with $P_0 = (1, 0, \ldots, 0)$. In practice, this result is available if n and x are not too large.
When the X_i are i.i.d. rv's with negative expectation, Dembo and Karlin ([DK92]) have investigated the asymptotic behaviour of H_n. More precisely they proved :

$$\lim_{n\to+\infty} \mathbb{P}\left(H_n \leq \frac{\ln n}{\lambda} + x\right) = \exp\left(-K^* \exp\left(-\lambda x\right)\right) \tag{1.1.1}$$

where K^* and λ depend only on the probability distribution of X_1.
When the X_i are i.i.d. with positive expectation, the growing of H_n is drastically different. The strong law of large numbers implies $S_n \underset{n\to+\infty}{\sim} \mathbb{E}(X_1)\, n$. Obviously $H_n = \max_{j\leq n} Y_j$ where $Y_j = S_j - \min_{i\leq j} S_i$. Since $\lim_{n\to+\infty} S_n = +\infty$ a.s.,then $-(\min_{j\leq n} S_j)$ converges a.s. to a finite r.v., when n goes to infinity. So $Y_j \underset{j\to+\infty}{\sim} S_j$ and $H_n \underset{n\to+\infty}{\sim} \mathbb{E}(X_1)\, n$. Consequently in this case, we can conclude that $\mathbb{E}(X_1)$ is the parameter which governs a phase transition phenomenon.

1.2 Here we investigate the case where $(X_i)_{i\geq 1}$ is a sequence of r.v's with null or "small" expectation.
We start with the centered case. We suppose that $(X_n)_{n\geq 1}$ is either a sequence of centered i.i.d. r.v's with variance $\sigma^2 > 0$ or a "good" Markov centered chain under its invariant probability with parameter σ (see the details in section 1.2). In this context, we prove that :

$$\frac{H_n}{\sqrt{n}} \underset{n\to\infty}{\xrightarrow{(d)}} \sigma\, B_1^*, \tag{1.1.2}$$

where $B_1^* = \max_{0\leq u\leq 1} |B_u|$, and $(B_u, u \geq 0)$ denotes a standard Brownian motion started at 0.
The distribution function of B_1^* is defined as a series (cf Proposition 8).

We generalize the previous case taking a family $\left\{(X_k^{(N)})_{k\geq 1}; N \geq 1\right\}$ of i.i.d. r.v's depending on a parameter N. We assume :

$$\lim_{N\to+\infty} \sqrt{N}\mathbb{E}\left(X_1^{(N)}\right) = \delta \in \mathbb{R}, \quad \lim_{N\to+\infty} Var\left(X_1^{(N)}\right) = \sigma^2 > 0. \tag{1.1.3}$$

If the sequence $(X_k)_{k\geq 1}$ does not depend on N, then (1.1.3) is equivalent to : $\mathbb{E}(X_1) = 0$ and $Var(X_1) = \sigma^2$. This means that this new setting generalizes the previous one.
Suppose from now that (1.1.3) holds.
Notice that $\mathbb{E}\left(X_1^{(N)}\right) \to 0$, when $N \to \infty$.
We prove (cf. proposition 11) that the analog of (1.1.2) is :

$$\frac{H_N^{(N)}}{\sqrt{N}} \xrightarrow[n\to\infty]{(d)} \sigma\xi_{\delta/\sigma}, \tag{1.1.4}$$

where $\xi_\gamma = \max_{0\leq u\leq 1}\{B(u) + \gamma u - \min_{0\leq s\leq u}(B(s) + \gamma s)\}$.

1.3 At this stage we would like to summarize different approximations of H_n, n going to infinity.

- If $\mathbb{E}(X_1) < 0$, following Dembo and Karlin ([DK92]) , the distribution of H_n is approximated by the law of $\frac{\ln n}{\lambda} + \eta$ where η is a r.v. whose distribution function is

$$\mathbb{P}(\eta \leq x) = \exp(-K^* \exp(-\lambda x)); \quad x \geq 0.$$

- If $\mathbb{E}(X_1) > 0$, H_n is a.s. equivalent to $\mathbb{E}(X_1)\, n$.

- If $\mathbb{E}(X_1) = 0$, the distribution of H_n can be estimated by the distribution of $(\sigma B_1^*)\sqrt{n}$.

- Suppose that X_1 has a finite variance σ^2 and $\mathbb{E}(X_1)$ is "small". This means that we can find n in a such way that $\sqrt{n}\mathbb{E}(X_1)$ is bounded by a constant say 10. We set $\delta = \sqrt{n}\mathbb{E}(X_1)$. This leads us to an approximation of the law of H_n by the distribution of $(\sigma\xi_{\delta/\sigma})\sqrt{n}$.

Tests making use of numerical simulations are led by [Da02] to compare the different methods of approximations.

1.4 The distribution function of ξ_γ is difficult to explicit. We prove that for any fixed $a > 0$, $\mathbb{P}(\xi_\gamma > a)$ is the sum of a series (cf. theorem 15).
Let us introduce

$$\xi_\gamma(t) = \max_{0\leq u\leq t}\left\{B(u) + \gamma u - \min_{0\leq s\leq u}(B(s) + \gamma s)\right\}, \quad t \geq 0. \tag{1.1.5}$$

and

$$T_a = \inf\left\{t \geq 0; B(t) + \gamma t - \min_{0\leq s\leq t}(B(s) + \gamma s) > a\right\}, \quad a > 0. \tag{1.1.6}$$

Obviously $\xi_\gamma = \xi_\gamma(1)$.
Taylor ([Tay75]) and Williams([Wil76]) have determined the Laplace transform of T_a :

$$\mathbb{E}\left[e^{-\lambda^2 T_a/2}\right] = \frac{\nu e^{\gamma a}}{\nu \cosh \nu a + \gamma \sinh \nu a}, \quad \lambda > 0. \tag{1.1.7}$$

where $\nu = \sqrt{\lambda^2 + \gamma^2}$.
The distribution of T_a and $(\xi_\gamma(t), t \geq 0)$ are linked by the relation :

$$\mathbb{P}\left(T_a < t\right) = \mathbb{P}\left(\xi_\gamma(t) > a\right), \quad \forall t \geq 0. \tag{1.1.8}$$

Suppose that α is a r.v. independent of $(B_t, t \geq 0)$ with exponential distribution, then :

$$\mathbb{P}\left(\xi_\gamma(\alpha) > a\right) = \mathbb{P}\left(T_a < \alpha\right) = \mathbb{E}\left[e^{-T_a}\right]; \quad \forall a > 0. \tag{1.1.9}$$

Consequently the distribution function of $\xi_\gamma(\alpha)$ is explicit :

$$1 - \mathbb{P}\left((\xi_\gamma(\alpha) \leq a\right) = \frac{\nu e^{\gamma a}}{\nu \cosh \nu a + \gamma \sinh \nu a}; \quad \forall a > 0. \tag{1.1.10}$$

1.5 As we said in 1.4, the distribution of ξ_γ is not easy to handle. So we investigate the tail of ξ_γ. We prove (cf Theorem 10) :

$$\mathbb{P}\left(\xi_\gamma \geq a\right) \underset{a\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}}\,\frac{1}{a}\,e^{-(\gamma-a)^2/2}. \tag{1.1.11}$$

We observe that $a \to \mathbb{P}\left(\xi_\gamma \geq a\right)$ goes slightly faster to 0, when $\gamma < 0$. This seems natural since $B_t + \gamma t$ goes to ∞ (resp. $-\infty$) when $\gamma > 0$ (resp.$\gamma < 0$) and $t \to \infty$.
These two different limit regimes illustrate the discussion initiated in 1.1 and 1.3.

1.6 Now let us briefly describe the organization of the paper. In section 1.2, we study the convergence of H_n when n goes to infinity, the underlying random variables X_i being centered. In section 1.3, we investigate the asymptotic behaviour of H_n when the X_i have a bias depending on N. The philosophy of this section is the following one : we give a short introduction to our approach. We state the results and detail only short proofs. The more technical proofs are postponed in section 1.4.

Acknowledgment. We would like to thank the referee for his interesting remarks and suggestions (in particular a direct proof of theorem 10).

1.2 Convergence of the local score in the centered case

Let $(X_n)_{n\geq 1}$ be a sequence of real valued random variables. $(S_k)_{k\geq 0}$ denotes the associated random walk :

$$S_0 = 0, \qquad S_k = \sum_{i=1}^{k} X_i; \qquad k \geq 1. \tag{1.2.1}$$

and the local score

$$H_n = \max_{0\leq i\leq j\leq n} (S_j - S_i) = \max_{0\leq i\leq j\leq n} (X_{i+1} + \ldots + X_j). \tag{1.2.2}$$

We define the sequence of score processes $\left(H^{(N)}\right)_{N\geq 1}$ which are piecewise linear processes :

$$\begin{cases} t \mapsto H^{(N)}(t) \text{ is linear on each interval of the form } \left[\frac{j}{N}; \frac{j+1}{N}\right] \\ H^{(N)}\left(\frac{j}{N}\right) = \frac{1}{\sqrt{N}} H_j. \end{cases} \tag{1.2.3}$$

In this section the sequence $(X_n)_{n\geq 1}$ will be either a sequence of i.i.d. centered variables with finite second moment or a stationary and irreducible Markov chain on a finite subset of $\mathbb{R}$. In the first case we set $\sigma^2 = Var(X_1)$, in the second one we suppose that $\mathbb{E}_\nu(X_1) = 0$ and

$$\sigma^2 = \mathbb{E}_\nu(X_1^2) + 2\sum_{k=2}^{\infty} \mathbb{E}_\nu(X_1 X_k), \tag{1.2.4}$$

where ν is the invariant distribution of $(X_n)_{n\geq 0}$.
σ^2 is well defined for the series (1.2.4) is convergent ([Bil68, p. 166]).
We are now able to state the main result of this section :

Theorem 7 *Let $(X_n)_{n\geq 1}$ be a sequence of random variables as above. Then the sequence of processes $\left(H^{(N)}(t),\ t \geq 0\right)$ converges in law to the process $(\sigma \max_{0\leq u\leq s} |B_u|, s \geq 0)$, as N tends to infinity.*

Proof : We just outline the proof, the complete developments are given in section 1.4.1.
Let $B^{(N)}$ be the piecewise linear process defined by

$$B^{(N)}\left(\frac{k}{N}\right) = \frac{1}{\sigma\sqrt{N}} S_k; \qquad k \geq 0. \tag{1.2.5}$$

and

$$t \mapsto B^{(N)}(t) \text{ is linear on each interval of the form } \left[\frac{k}{N}, \frac{k+1}{N}\right] \tag{1.2.6}$$

It is well known ([Bil68]) that $\left(B^{(N)}(s), s \geq 0\right)$ converges to the standard Brownian motion. We easily check that $\left(H^{(N)}(s), s \geq 0\right)$ may be approached by a continuous function of $\left(B^{(N)}(s), s \geq 0\right)$ up to a remainder term R_N which converges to 0. This completes the proof of theorem 7. □

An important application of theorem 7 is the convergence of the local score :

Proposition 8 *1)* $\dfrac{H_n}{\sqrt{n}}$ *converges in distribution, as* $n \to \infty$, *to* $\sigma\, B_1^*$, *where* $B_1^* = \max_{0 \leq u \leq 1} (|B_u|)$.
2) The cumulative distribution function (c.d.f) of B_1^* *is :*

$$\mathbb{P}(B_1^* \leq x) = \frac{2}{\pi} \sum_{k \in \mathbb{Z}} \frac{(-1)^k}{2k+1} \exp\left(-\frac{(2k+1)^2 \pi^2}{8x^2}\right), \quad x \geq 0. \tag{1.2.7}$$

Proof : Theorem 7 implies the convergence in law of the random variable $\frac{H_n}{\sqrt{n}}$. Indeed

$$\frac{H_n}{\sqrt{n}} = H^{(n)}\left(\frac{n}{n}\right) = H^{(n)}(1), \qquad \forall\, n \geq 0.$$

The equality (1.2.7) is classical and may be deduced from [BS96] (p.146) see 1.4.6 by a short calculus. □

Remark 9 Theorem 7 implies the convergence of $T_a(H)/a^2$, as a tends to infinity, where $T_a(H) = \inf\{k \geq 0;\ H_k > a\}$, $a > 0$. Given $a \in \mathbb{R}^+$, then

$$\frac{T_a(H)}{a^2} \underset{a \to \infty}{\xrightarrow{(d)}} \frac{1}{\sigma^2 (B_1^*)^2}. \tag{1.2.8}$$

Proof : H_k is a non decreasing process, so :

$$\left\{\frac{T_{x\sqrt{N}}(H)}{N} < t\right\} \subset \left\{\frac{H_{[Nt]}}{\sqrt{N}} > x\right\}$$

and

$$\left\{\frac{H_{[Nt]-1}}{\sqrt{N}} > x\right\} \subset \left\{\frac{T_{x\sqrt{N}}(H)}{N} < t\right\}.$$

We know also that $\dfrac{H_{[Nt]-1}}{\sqrt{N}}$ and $\dfrac{H_{[Nt]}}{\sqrt{N}}$ have the same limit : $\sigma\sqrt{t} B_1^*$. Then

$$\mathbb{P}\left(\frac{T_{x\sqrt{N}}}{N} < t\right) \underset{N \to \infty}{\longrightarrow} \mathbb{P}\left(\sigma\sqrt{t} B_1^* > x\right) = \mathbb{P}\left(\frac{x^2}{\sigma^2 (B_1^*)^2} < t\right). \tag{1.2.9}$$

Let $a = x\sqrt{N}$, (1.2.8) follows immediately. □

1.3 Convergence in the non-centered case.

3.1 $(B_t; t \geq 0)$ will denote as previously, a standard Brownian motion starting at 0. In this section we suppose that $(X_n)_{n\geq 0}$ is a sequence of i.i.d random variables and that the law of X_1 depends upon N, N being the order of approximation. More precisely, we assume :

$$\lim_{N\to\infty} Var(X_1) = \sigma^2 > 0 \quad ; \quad \lim_{N\to\infty} \sqrt{N}\mathbb{E}(X_1) = \delta \in \mathbb{R}. \tag{1.3.1}$$

In this setting it is easy to prove (cf proposition 11) that $H_N/\sqrt{N}$ converges in distribution, when N goes to infinity, to $\sigma\xi_{\delta/\sigma}$, where

$$\xi_\gamma = \max_{0\leq u\leq 1}\left\{B(u) + \gamma u - \min_{0\leq s\leq u}(B(s) + \gamma s)\right\}. \tag{1.3.2}$$

In the sequel we focus on the law of ξ_γ. It is convenient to introduce :

$$\phi^{(\gamma)}(a) = e^{-\gamma\, a}\mathbb{P}(\xi_\gamma > a), \quad a \geq 0. \tag{1.3.3}$$

Let us briefly detail our approach. We state the main result (theorem 10) at the end of the subsection.
In section 1.2 we have determined the distribution of ξ_γ when $\gamma = 0$. This brings us to remove the drift term, using Girsanov's transformation. Using moreover pathwise properties of Brownian motion we prove (cf proposition 12 and theorem 13) :

$$\phi^{(\gamma)}(a) = \frac{1}{a}\int_{[0,+\infty[^2} \mathbb{1}_{\{u\leq 1\}}\, \exp\left\{-\gamma t - \gamma^2\, u/2\right\} \mu_a(u)\, F_t^{(\gamma)}(1-u, 1/a)\, \mathrm{d}u\, \mathrm{d}t \tag{1.3.4}$$

where $F_t^{(\gamma)}$ can be expressed as an expectation of a positive r.v. :

$$F_t^{(\gamma)}(x,b) = \mathbb{E}\left(\mathbb{1}_{\left\{0\leq\tau_t\leq x,\ 0\leq B^*_{\tau_t}\leq 1/b\right\}} e^{-\gamma^2\tau_t/2}\right); \qquad x \geq 0,\ b \geq 0,\ t \geq 0. \tag{1.3.5}$$

The two random variables τ_t and $B^*_{\tau_t}$ are defined as follows :

- τ_t is the first time where the local time at 0 of Brownian motion $(B_u, u \geq 0)$ reaches level t,
- $(B^*_t, t \geq 0)$ is the process : $B^*_t = \sup_{0\leq u\leq t}|B_u|$.

For any positive number a, the function μ_a is known (cf (1.3.16) and (1.3.17)).
This leads us to determine the joint distribution of $(\tau_t, B^*_{\tau_t})$.
The decomposition of the Brownian path up to time τ_t (namely $(B_u; 0 \leq u \leq \tau_t)$), conditionally to $B^*_{\tau_t}$ leads to some recursive structure. This generates two analytic counterparts.

- The density function θ_t of $(\tau_t, B^*_{\tau_t})$ satisfies an integral equation (proposition 16),

- $F_t^{(\gamma)}$ is solution of an integral equation (cf (1.3.20)).

Moreover relation (1.3.20) yields to express $F_t^{(\gamma)}$ as sum of a series (Theorem 15). Unfortunately the coefficients are not explicit and are determined by a recursive algorithm.
However relation (1.3.20) is rich enough since we determine the decay rate of $a \to P(\xi_\gamma > a)$, $a \to \infty$. More precisely

Theorem 10 *For all γ in $\mathbb{R}$:*

$$\mathbb{P}(\xi_\gamma \geq a) \underset{a\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}}\, e^{-\gamma^2/2}\, \frac{1}{a}\, e^{\gamma a - a^2/2} = 2\sqrt{\frac{2}{\pi}}\, \frac{1}{a} e^{-(\gamma-a)^2/2}. \qquad (1.3.6)$$

Two proofs of Theorem 10 will be given. The first one is a consequence of Theorem 13 and is postponed in section 1.4.4. The second one suggested by the referee will be developed at the end of this section.

3.2 We now prove the main result mentioned in 3.1 (Theorems 13 and 15). To help the reader we restrict ourself to short and easy proofs, the more technical points are postponed in the last section 1.4.

Recall that $(X_n)_{n\geq 0}$ will denote a sequence of i.i.d. random variables such that the law of X_1 depends upon a parameter N. We suppose that (1.3.1) holds. For instance, we can choose

$$\mathbb{P}(X_i = 1) = p_N = \frac{1}{2} + \frac{\delta}{2\sqrt{N}} \text{ and } \mathbb{P}(X_i = -1) = q_N = \frac{1}{2} - \frac{\delta}{2\sqrt{N}},$$

for N large enough so that $|\frac{\delta}{\sqrt{N}}| < 1$. Then

$$\mathbb{E}(X_1) = p_N - q_N = \frac{\delta}{\sqrt{N}} \quad \text{and} \quad Var(X_1) = 1 - \frac{\delta^2}{N}.$$

We set $a_N = \mathbb{E}(X_1)$. Define $B^{(N)}$ as

$$B^{(N)}\left(\frac{k}{N}\right) = \frac{1}{\sigma\sqrt{N}}\,(S_k - \mathbb{E}(S_k)) = \frac{1}{\sigma\sqrt{N}}\,(S_k - k\,a_N)\,; \qquad k \geq 0. \qquad (1.3.7)$$

and

$$t \mapsto B^{(N)}(t) \text{ is linear on each interval of the form } \left[\frac{k}{N}, \frac{k+1}{N}\right] \qquad (1.3.8)$$

The process $(H^{(N)}(t), t \geq 0)$ is defined by the same procedure as in the centered case, i.e. formula (1.2.3). It can be shown ([Bil68], p.68) that $(B^{(N)}(t), t \geq 0)$ converges in distribution to $(B(t), t \geq 0)$. $(H^{(N)}(t), t \geq 0)$

is a continuous functional of $(B^{(N)}(t), t \geq 0)$, this implies the convergence of $H_{[Nt]}/\sqrt{N}$.

Proposition 11 *1. Let $t > 0$. As N tends to ∞,*

$$\frac{H_{[Nt]}}{\sqrt{N}} = \frac{1}{\sqrt{N}} \max_{1 \leq i \leq j \leq [Nt]} (S_j - S_i) \xrightarrow{(d)} \sigma \xi_{\delta/\sigma}(t),$$

where

$$\xi_\gamma(t) = \max_{0 \leq u \leq t} \left\{ B(u) + \gamma u - \min_{0 \leq s \leq u} (B(s) + \gamma s) \right\}. \tag{1.3.9}$$

2. In particular $H_n/\sqrt{n}$ converges in distribution, as $n \to \infty$, to $\sigma \xi_{\delta/\sigma}$, where $\xi_\gamma = \xi_\gamma(1)$.

Proof : (see section 1.4.2 for a complete proof). □

Remark : The classical scaling property of Brownian motion (i.e. $(B_s; s \geq 0) \overset{(d)}{=} (\sqrt{t} B_{s/t}; s \geq 0)$, for any $t > 0$) implies that :

$$\xi_\gamma(t) \overset{(d)}{=} \sqrt{t} \xi_{\gamma\sqrt{t}}, \quad \text{for any } t > 0. \tag{1.3.10}$$

This leads us to determine the distribution of ξ_γ.

Proposition 12 *For all $a > 0$ and $\gamma \in \mathbb{R}$, we set*

$$\phi^{(\gamma)}(a) = e^{-\gamma a} \mathbb{P}(\xi_\gamma > a). \tag{1.3.11}$$

Then

$$\phi^{(\gamma)}(a) = \mathbb{E}\left[\mathbb{1}_{\{\tau_Z + T_a < 1\}} \exp\left\{ -\gamma Z - \frac{\gamma^2}{2}(\tau_Z + T_a) \right\} | B^*_{\tau_Z} < a \right], \quad \gamma \in \mathbb{R}, \tag{1.3.12}$$

where

- *τ_t denotes the first time where the local time at 0 of Brownian motion $(B_t; t \geq 0)$ reaches t,*
- *T_a is the first time where a Bessel process of dimension 3, starting at 0, hits a,*
- *Z is a random exponential variable of parameter a (i.e. its density function is $\frac{1}{a} e^{-x/a} \mathbb{1}_{\{x>0\}}$).*
- *$(B^*_u; u \geq 0)$ is the process : $B^*_u = \sup_{0 \leq s \leq u} |B_s|$, $u \geq 0$.*

- *for any $a > 0$, $(B_t; t \geq 0)$, Z and T_a are independent.*

Proof : We make use on one hand Girsanov's transformation to reduce to the Brownian case and on second hand some sample path properties. See section 1.4.3. □

A priori we only need to handle $\phi^{(\gamma)}$. However $\phi^{(\gamma)}$ coincides with $\phi_\gamma^{(\gamma)}$, the function $\phi_\lambda^{(\gamma)}$ being defined as follows :

$$\phi_\lambda^{(\gamma)}(a) = \mathbb{E}\left[\mathbb{1}_{\{\tau_Z + T_a < 1\}} \exp\left\{-\gamma Z - \frac{\lambda^2}{2}(\tau_Z + T_a)\right\} | B^*_{\tau_Z} < a\right], \quad \lambda \in \mathbb{R}, \tag{1.3.13}$$

In our approach it is not more difficult to deal with $\phi_\lambda^{(\gamma)}$ instead of $\phi^{(\gamma)}$. Formula (1.3.12) gives a simple stochastic interpretation of $\phi_\lambda^{(\gamma)}$, but we have to express $\phi_\lambda^{(\gamma)}$ under a more convenient form for computation purpose. The analytic transcription of (1.3.12) is the following :

Theorem 13 *Let $\lambda \in \mathbb{R}$ be fixed, then for any $a > 0$*

$$\phi_\lambda^{(\gamma)}(a) = \frac{1}{a} \int_{[0,+\infty[^2} \mathbb{1}_{\{u \leq 1\}} \exp\left\{-\gamma t - \lambda^2 u/2\right\} \mu_a(u) F_t^{(\lambda)}(1-u, 1/a) \, \mathrm{d}u \, \mathrm{d}t \tag{1.3.14}$$

where

$$F_t^{(\lambda)}(x,b) = \mathbb{E}\left(\mathbb{1}_{\left\{0 \leq \tau_t \leq x,\ 0 \leq B^*_{\tau_t} \leq 1/b\right\}} e^{-\lambda^2 \tau_t/2}\right); \qquad x \geq 0,\ b \geq 0,\ t \geq 0, \tag{1.3.15}$$

and μ_a is the density function of T_a :

$$\mu_a(t) = \frac{1}{a^2} \mu_1\left(\frac{t}{a^2}\right), \tag{1.3.16}$$

and

$$\mu_1(t) = \frac{1}{\sqrt{2\pi}\, t^{3/2}} \sum_{k \in \mathbb{Z}} \left(-1 + \frac{(1+2k)^2}{t}\right) \exp - \frac{(1+2k)^2}{2t}. \tag{1.3.17}$$

Furthermore μ_1 may be expressed as ([BPY01], p.8 and 24) :

$$\mu_1(t) = \frac{d}{dt} \sum_{n=-\infty}^{\infty} (-1)^n e^{-(n^2\pi^2 t)/2}. \tag{1.3.18}$$

Remark 14 Let us define $T_a^* = \inf\{t > 0, |B_t| > a\}$, then $L^0_{T_a^*}$ is an exponential random variable of parameter a.
Since $\left\{B^*_{\tau_t} < a\right\} = \left\{L^0_{T_a^*} > t\right\}$, obviously $\mathbb{P}\left(B^*_{\tau_t} < a\right) = e^{-t/a}$.

Proof of theorem 13 : The random variables involved in equation (1.3.12) being independent, we have :

$$\phi_\lambda^{(\gamma)}(a) = \frac{1}{a}\int_{[0,+\infty[^2} \mathbb{E}\left[\mathbb{1}_{\{\tau_t+u<1\}} e^{\left(-\gamma\, t - \frac{\lambda^2}{2}(\tau_t+u)\right)} \mid B^*_{\tau_t} < a\right] \mu_a(u) e^{-t/a}\,\mathrm{d}u\,\mathrm{d}t, \tag{1.3.19}$$

where μ_a denotes the density function of T_a.
Using Remark 14, equation (1.3.14) follows immediately. □

We focus our attention on $F_t^{(\lambda)}$. The decomposition of the Brownian path $(B_u, 0 \le u \le \tau_t)$, conditionally to $B^*_{\tau_t}$ leads to some recursive structure. This has an analytic consequence : $F_t^{(\lambda)}$ is solution of an integral equation.

Theorem 15 *Let $\lambda \in \mathbb{R}$ and $t \ge 0$ be two fixed parameters.*

1. *$F^{(\lambda)}$ satisfies the integral equation :*

$$F_t^{(\lambda)}(x,a) = F_t^{(\lambda)}(x,0) - t\left(A^{(\lambda)}F_t^{(\lambda)}\right)(x,a), \quad (x,a)\in\mathbb{R}_+^2, \tag{1.3.20}$$

with

$$\left(A^{(\lambda)}\psi\right)(x,a) = \int_{[0,+\infty[^2} \mathbb{1}_{\{u\le a,\, y\le x\}} \mu^{(2)}_{1/u}(y)\, e^{-\lambda^2 y/2}\psi(x-y,u)\,\mathrm{d}y\,\mathrm{d}u, \tag{1.3.21}$$

$$F_t^{(\lambda)}(x,0) = \frac{t}{\sqrt{2\pi}}\int_0^x \exp\left(-\frac{\lambda^2 z}{2} - \frac{t^2}{2z}\right)\frac{\mathrm{d}z}{z^{3/2}}, \tag{1.3.22}$$

and $\mu_a^{(2)}(u) = (\mu_a * \mu_a)(u) = \dfrac{1}{a^2}\mu_1^{(2)}(u/a^2)$.

Recall (cf [BPY01]) that :

$$\mu_1^{(2)}(t) = \frac{d}{dt}\left(\frac{8\sqrt{2}}{\sqrt{\pi}t^{3/2}}\sum_{n=1}^{+\infty} n^2 e^{-2n^2/t}\right). \tag{1.3.23}$$

2. *Furthermore $F_t^{(\lambda)}$ can be expressed as a series :*

$$F_t^{(\lambda)}(x,a) = \sum_{k=0}^{+\infty}(-1)^k\, t^k\, \alpha_t^{(k)}(x,a), \tag{1.3.24}$$

where

$$\alpha_t^{(0)}(x,a) = F_t^{(\lambda)}(x,0), \tag{1.3.25}$$

$$\alpha_t^{(k+1)}(x,a) = \left(A^{(\lambda)}\alpha_t^{(k)}\right)(x,a). \tag{1.3.26}$$

The convergence of (1.3.24) holds uniformly for $(x,a)\in\mathbb{R}_+\times[0,M]$, for any $M\ge 0$.

3. For $\gamma \in \mathbb{R}$ and $a > 0$, $\mathbb{P}(\xi_\gamma > a)$ admits the following development

$$\mathbb{P}(\xi_\gamma > a) = \frac{e^{\gamma a}}{a} \sum_{k\geq 0} (-1)^k \int_{[0,+\infty[^2} \mathbb{1}_{\{u\leq 1\}} e^{-\gamma t - \lambda^2 u/2} \mu_a(u) t^k \alpha_t^{(k)}(1-u, 1/a) \mathrm{d}u\, \mathrm{d}t. \tag{1.3.27}$$

Proof : see section 1.4.5. □

Making use of Theorem 15 (especially formula (1.3.20)) we prove that the two dimensional random variable $(\tau_t, B^*_{\tau_t})$ admits for any $t > 0$ a density function θ_t and it verifies an integral equation (1.3.28). As we notice in 3.1, θ_t is unknown, therefore (1.3.28) is interesting.
As formula (1.3.14) shows, $\phi_\lambda^{(\gamma)}$ can be written as an integral of an explicit function of four variables (u, t, x, y) with respect to the positive measure on $\mathbb{R}_+^4$: $\theta_t(x, y)\, \mathrm{d}u\, \mathrm{d}t\, \mathrm{d}x\, \mathrm{d}y$. However this formula is in practice unusable. In particular the asymptotic development (1.3.24) of $F_t^{(\lambda)}$ cannot be deduced from it. This justifies our choice : $F_t^{(\lambda)}$ is the right parameter.

Proposition 16 *Let $t > 0$. The random variable $(\tau_t, B^*_{\tau_t})$ has a density function θ_t. Moreover θ_t verifies :*

$$\theta_t(x, a) = \frac{t}{a^2} \int_{[0,+\infty[^2} \mathbb{1}_{[0,x]\times[0,a]}(y, b) \mu_a^{(2)}(x - y) \theta_t(y, b)\, \mathrm{d}y\, \mathrm{d}b. \tag{1.3.28}$$

Proof : Let f be the distribution of $(\tau_t, B^*_{\tau_t})$.
Then $f([0, x] \times [0, a]) = F^{(0)}(x, 1/a)$. We choose $\lambda = 0$ and replace a by $1/a$ in equation (1.3.20), we get :

$$f([0, x] \times [a, +\infty[) = t \int_0^{1/a} \mathrm{d}u \left(\int_0^x \mu_{1/u}^{(2)}(y) f([0, x - y] \times [0, 1/u]) \mathrm{d}y \right). \tag{1.3.29}$$

Let η_v be the positive measure $\eta_v(\mathrm{d}y) = \mathbb{1}_{\{y>0\}} \hat{\eta}_v(y) \mathrm{d}y$, where $\hat{\eta}_v(y) = \mu_v^{(2)}(y) \mathbb{1}_{\{y>0\}}$.
But

$$\begin{aligned} (f(., [0, v]) * \eta_v)([0, x]) &= \int_0^x \mu_v^{(2)}(y) f([0, x - y] \times [0, v]) \mathrm{d}y, \\ &= \int_0^x \{(f(., [0, v]) * \hat{\eta}_v)(y) \mathrm{d}y. \end{aligned}$$

The new relation obtained by setting $v = 1/a$ in (1.3.29) implies that $(\tau_t, B^*_{\tau_t})$ has a density θ_t and

$$\theta_t(x, a) = \frac{t}{a^2} \int_0^x \mu_a^{(2)}(x - y) f(\mathrm{d}y, [0, a]),$$

$$= \frac{t}{a^2}\int_{[0,+\infty[^2} \mathbb{1}_{[0,x]\times[0,a]}(y,b)\mu_a^{(2)}(x-y)\theta_t(y,b)\mathrm{d}b\,\mathrm{d}y.$$

□

To end up this section we give a direct proof of Theorem 10 suggested by the referee.

Proof of Theorem 10 : Let us introduce some notations. We state :

$$X_s = B_s + \gamma s, \quad I_t = \inf_{0\le s\le t} X_s, \quad Y_t = X_t - I_t, \quad T_a = \inf\{t \ge 0 : Y_t = a\}.$$

Then

$$\mathbb{P}(\xi_\gamma > a) = \mathbb{P}(T_a < 1). \tag{1.3.30}$$

Williams([Wil76]) and Taylor ([Tay75]) have determined the Laplace transform of T_a :

$$\mathbb{E}\left[e^{-\frac{\lambda^2}{2}T_a}\right] = \frac{\nu e^{\gamma a}}{\nu\cosh\nu a + \gamma\sinh\nu a}. \tag{1.3.31}$$

where $\nu = \sqrt{\lambda^2+\gamma^2}$.
We are able to invert this formula (cf. step 1 below), i.e. to determine the density function of T_a. Then using (1.3.30), we explicit the asymptotic behaviour of $\mathbb{P}(\xi_\gamma > a)$, $a \to \infty$.
1) By an easy computation we have :

$$\begin{aligned}\mathbb{E}\left[e^{-\frac{\lambda^2}{2}T_a}\right] &= \frac{2\nu e^{(\gamma-\nu)a}}{\gamma+\nu}\frac{1}{1+\frac{\nu-\gamma}{\gamma+\nu}e^{-2\nu a}},\\ &= 2\nu e^{\gamma a}\left(\sum_{k\ge0}(-1)^k\frac{(\nu-\gamma)^k}{(\gamma+\nu)^{k+1}}e^{-(2k+1)\nu a}\right).\end{aligned}$$

Let L_k be the Laguerre polynomial of order k ([Wil76], p.168). Its Laplace transform is known ([Wil76], (7) p.170) :

$$\int_0^{+\infty} e^{-sx}L_k(x)\mathrm{d}x = \frac{(s-1)^k}{s^{k+1}}; \quad s \ge 0.$$

This yields to

$$\mathbb{E}\left[e^{-\frac{\lambda^2}{2}T_a}\right] = 2e^{\gamma a}\left(\sum_{k\ge0}(-1)^k\int_0^{+\infty}\nu e^{-\nu(t+(2k+1)a)-\gamma t}L_k(2\gamma t)\mathrm{d}t\right). \tag{1.3.32}$$

Let us recall the integral representation of K_ρ ([Wat95] (15), p.183) :

$$K_\rho(z) = \frac{1}{2}\left(\frac{z}{2}\right)^\rho\int_0^{+\infty}\frac{1}{y^{\rho+1}}e^{-(y+z^2/4y)}\mathrm{d}y.$$

But $K_{1/2}$ and $K_{3/2}$ are explicit ([Wat95], (12), (13) p.80) :

$$K_{1/2}(z) = \sqrt{\frac{\pi}{2z}}e^{-z}; \quad K_{3/2}(z) = \sqrt{\frac{\pi}{2z}}e^{-z}\left(1+\frac{1}{z}\right).$$

In particular

$$\rho e^{-\rho(t+(2k+1)a)} = \frac{1}{\sqrt{2\pi}}\int_0^{+\infty} e^{-\lambda^2 x/2}\frac{(t+(2k+1)a)^2 - x}{x^{5/2}}e^{-(\gamma^2 x+(t+(2k+1)a)^2/x)/2}\mathrm{d}x.$$

Consequently we are able to invert (1.3.32) : T_a has a density ϕ_a and

$$\phi_a(x) = \sqrt{\frac{2}{\pi}}e^{\gamma a}\frac{e^{-\gamma^2 x/2}}{x^{5/2}}\left(\sum_{k\geq 0}\psi_{a,k}(x)\right), \tag{1.3.33}$$

where

$$\psi_{a,k}(x) = (-1)^k\int_0^{+\infty}\left((t+(2k+1)a)^2 - x\right)e^{-\gamma t}L_k(2\gamma t)e^{-(t+(2k+1)a)^2/2x}\mathrm{d}t. \tag{1.3.34}$$

2) We say that $h_a^1(x)$ is uniformly equivalent to $h_a^2(x)$, as $a\to\infty$, x belonging to $[0;1]$, if

$$\lim_{a\to\infty}\left(\sup_{x\in[0;1]}\frac{h_a^1(x)}{h_a^2(x)}\right) = 1.$$

We write $h_a^1(x) \underset{a\to\infty}{\overset{u}{\sim}} h_a^2(x)$.
We prove :

$$\psi_{a,0}(x) \underset{a\to\infty}{\overset{u}{\sim}} xae^{-a^2/2x}. \tag{1.3.35}$$

We set $t+a = \sqrt{x}u$ in (1.3.34) (with $k=0$) :

$$\psi_{a,0}(x) = x^{3/2}e^{\gamma a}\int_{a/\sqrt{x}}^{+\infty}(u^2-1)e^{-u^2/2}e^{-\gamma\sqrt{x}u}\mathrm{d}u. \tag{1.3.36}$$

But

$$\left(-ue^{-u^2/2}\right)' = (u^2-1)e^{-u^2/2}, \tag{1.3.37}$$

then integrating by part in (1.3.36) we obtain :

$$\psi_{a,0}(x) = x^{3/2}e^{\gamma a}\left(\frac{a}{\sqrt{x}}e^{-a^2/2x}e^{-\gamma a} - \gamma\sqrt{x}\int_{a/\sqrt{x}}^{+\infty}ue^{-u^2/2}e^{-\gamma\sqrt{x}u}\mathrm{d}u\right).$$

Since $u \leq \frac{\sqrt{x}}{a}u^2$ for any $u\in[a/\sqrt{x};+\infty[$ and $x\in[0;1]$

$$\frac{xae^{-a^2/2x}}{1+\gamma/a} \leq \psi_{a,0}(x) \leq xae^{-a^2/2x}.$$

(1.3.35) follows immediately.

3) We claim that

$$\sum_{k\geq 0} \psi_{a,k}(x) \underset{a\to\infty}{\overset{u}{\sim}} \psi_{a,0}(x). \tag{1.3.38}$$

Suppose $k \geq 1$, $a \geq 1$ and $x \in [0;1]$.
Then

$$(t+(2k+1)a)^2 \geq a^2 \geq 1 > x; \quad \forall t \geq 0. \tag{1.3.39}$$

Recall ([Wid41] theorem 17a p.168) :

$$\mid L_k(x) \mid \leq e^{x/2}; \quad x \in \mathbb{R}. \tag{1.3.40}$$

Setting $t+(2k+1)a = \sqrt{x}u$, we obtain :

$$\mid \psi_{a,k}(x) \mid \leq x^{3/2} \int_{(2k+1)a/\sqrt{x}}^{+\infty} (u^2-1)e^{-u^2/2}\mathrm{d}u.$$

By (1.3.37) the integral can be computed explicitly :

$$\mid \psi_{a,k}(x) \mid \leq (2k+1)xae^{-(2k+1)^2a^2/2x}.$$

But $x \in]0;1[$, then

$$\mid \psi_{a,k}(x) \mid \leq \left(x\,a\,e^{-a^2/2x}\right)\left((2k+1)e^{-(2k^2+2k)a^2}\right).$$

Since $k \geq 1$ and $a > 1$,

$$\mid \psi_{a,k}(x) \mid \leq \left(x\,a\,e^{-a^2/2x}\right)\left((2k+1)e^{-2k^2}e^{-2a^2}\right). \tag{1.3.41}$$

This demonstrates (1.3.38).

4) Let us en the proof of Theorem 10. Using both (1.3.30), (1.3.33), (1.3.35) and (1.3.38) we have :

$$\mathbb{P}\left(\xi_\gamma \geq a\right) \underset{a\to\infty}{\sim} \sqrt{\frac{2}{\pi}}ae^{\gamma a}I(a),$$

where

$$I(a) = \int_0^1 \frac{1}{x^{3/2}}e^{-\gamma^2 x/2}e^{-a^2/2x}\mathrm{d}x.$$

We set $x = a^2/(a^2+y)$, we obtain

$$I(a) = \frac{e^{-a^2/2}}{a^2}\int_0^{+\infty}\sqrt{\frac{a^2}{a^2+y}}e^{-\gamma^2a^2/2(a^2+y)}e^{-y/2}\mathrm{d}y.$$

Consequently

$$I(a) \underset{a\to\infty}{\sim} 2\frac{e^{-(\gamma^2+a^2)/2}}{a^2}.$$

This proves Theorem 10. □

1.4 Technical proofs

This section is devoted to the proofs of Theorems 7, 15, 10, Propositions 11, 12 and formulae (1.2.7), (1.3.17), (1.3.18) and (1.3.23).

1.4.1 Proof of Theorem 7

Let $(X_n)_{n\geq 1}$ be a sequence of r.v.'s and $(S_n)_{n\geq 0}$ the random walk :

$$S_0 = 0, \qquad S_n = \sum_{k=1}^{n} X_k, \quad n \geq 1.$$

We consider two cases :
a)(X_n) are i.i.d. centered random variables with finite second moment and $\sigma^2 = Var(X_1)$.
b)(X_n) is an irreducible Markov chain taking its values in a finite subset of $\mathbb{R}$. We denote by ν its invariant distribution. σ^2 is the parameter defined by (1.2.4).
Given an integer $N \geq 0$, we consider the piecewise linear process $B^{(N)}(t)$

$$\begin{cases} B^{(N)}\left(\frac{k}{N}\right) = \frac{1}{\sigma\sqrt{N}}\left(S_k - \mathbb{E}(S_k)\right) = \frac{1}{\sigma\sqrt{N}}\, S_k; \qquad k \geq 0, \\ t \mapsto B^{(N)}(t) \text{ is linear on each interval of the form } \left[\frac{k}{N}, \frac{k+1}{N}\right]. \end{cases} \tag{1.4.1}$$

Our approach is based on the two following results.

Theorem 17 (Billingsley, [Bil68], p. 68 and [Bil68], p.166 and p.174)
The process $\left(B^{(N)}(t), t \geq 0\right)$ *converges in law, as* N *tends to* $+\infty$*, to the standard linear Brownian motion* $(B(t), t \geq 0)$.

Theorem 18 (Skorokhod's theorem ([IW81], p. 9)) *Let* (S, γ) *be a complete separable metric space,* P *and* P_n*,* $n = 1, 2, \ldots$ *be probability measures on* $(S, \mathcal{B}(S))$ *so that* $P_n \underset{N\to\infty}{\Longrightarrow} P$*. Then, we can construct, on a probability space* $(\underline{\Omega}, \underline{\mathcal{B}}, \underline{P})$*, S-valued random variables* X_n*,* $n = 1, 2, \ldots$ *and* X *such that*

1. $P_n = \mathcal{L}(X_n)$*,* $n = 1, 2, \ldots$ *and* $P = \mathcal{L}(X)$
2. X_n *converges to* X *almost everywhere.*

Proof of Theorem 7: The proof is divided into two steps.
Recall that

$$H_k = \max_{0\leq i\leq j\leq k} (S_j - S_i); \quad k \geq 0.$$

Let us introduce the linear interpolation of $(H_k)_{k\geq 0}$. This function $\left(H^{(N)}(t); t \geq 0\right)$ depending on the parameter N is defined as follows :

$$H^{(N)}(t) = \frac{1}{\sqrt{N}}\left\{H_{[Nt]} + (Nt - [Nt])\left(H_{[Nt]+1} - H_{[Nt]}\right)\right\}, \quad t \geq 0. \tag{1.4.2}$$

1. Relation (1.4.1) implies :

$$S_k = \sigma\sqrt{N}\, B^{(N)}\left(\frac{k}{N}\right). \tag{1.4.3}$$

Then

$$\begin{aligned} H_{[N\,t]} &= \sigma\sqrt{N} \max_{0\le i\le j\le [N\,t]} \left\{ B^{(N)}\left(\frac{j}{N}\right) - B^{(N)}\left(\frac{i}{N}\right)\right\}, \\ &= \sigma\sqrt{N} \max_{0\le \frac{i}{N}\le \frac{j}{N}\le \frac{[N\,t]}{N}} \left\{ B^{(N)}\left(\frac{j}{N}\right) - B^{(N)}\left(\frac{i}{N}\right)\right\}. \end{aligned}$$

$B^{(N)}$ being piecewise linear, then the maximum on $\{0 \le \frac{i}{N} \le \frac{j}{N} \le \frac{[N\,t]}{N}\}$ is equal to the maximum on $\{0 \le u \le v \le \frac{[N\,t]}{N}\}$ and

$$H_{[N\,t]} = \sigma\sqrt{N} \max_{0\le u\le v\le \frac{[N\,t]}{N}} \left\{ B^{(N)}(v) - B^{(N)}(u)\right\}.$$

Finally $H^{(N)}(t)$ can be written as follows :

$$H^{(N)}(t) = \sigma\left(\max_{0\le u\le v\le \frac{[N\,t]}{N}} \left\{ B^{(N)}(v) - B^{(N)}(u)\right\} + R_N(t)\right), \tag{1.4.4}$$

$$\begin{aligned} R_N(t) = (N\,t - [N\,t]) \Bigg(& \max_{0\le u\le v\le \frac{[N\,t]+1}{N}} \left\{ B^{(N)}(v) - B^{(N)}(u)\right\} \\ & - \max_{0\le u\le v\le \frac{[N\,t]}{N}} \left\{ B^{(N)}(v) - B^{(N)}(u)\right\}\Bigg). \end{aligned} \tag{1.4.5}$$

2. We apply theorems 17 and 18 with $S = \mathcal{C}([0,T],\mathbb{R})$ and γ the Wiener measure, T being fixed. Then there exist $(\underline{\Omega}, \underline{\mathcal{B}}, \underline{P})$, $\underline{B}^{(N)}$ and $\underline{B}$ such that $\underline{B}^{(N)}$ converges almost surely to $\underline{B}$ a standard Brownian motion on $(\underline{\Omega}, \underline{\mathcal{B}}, \underline{P})$. Let $\underline{R}_N$ (resp. $\underline{H}^{(N)}$) be the process defined by (1.4.5) (resp. (1.4.4)) where $B^{(N)}$ is replaced by $\underline{B}^{(N)}$.
But $B^{(N)}$ and $\underline{B}^{(N)}$ have the same law, then :

$$\left(H^{(N)}(t), t\ge 0\right) \overset{(d)}{=} \left(\underline{H}^{(N)}(t), t\ge 0\right).$$

If we prove that $\underline{H}^{(N)}$ converge a.s., then the previous identity implies the convergence in distribution of $H^{(N)}$.
Since the convergence of $\underline{B}^{(N)}$ holds in the space of continuous functions, for any $t \in [0,T]$:

$$\max_{0\le u\le v\le \frac{[N\,t]+1}{N}} \left\{ \underline{B}^{(N)}(v) - \underline{B}^{(N)}(u)\right\} \underset{N\to\infty}{\xrightarrow{a.s}} \max_{0\le u\le v\le t} \left\{\underline{B}(v) - \underline{B}(u)\right\},$$

$$\max_{0\le u\le v\le \frac{[N\,t]}{N}} \left\{ \underline{B}^{(N)}(v) - \underline{B}^{(N)}(u) \right\} \xrightarrow[N\to\infty]{a.s} \max_{0\le u\le v\le t} \{\underline{B}(v) - \underline{B}(u)\}\,.$$

Moreover as $0 \le N\,t - [N\,t] \le 1$, then

$$\underline{R}_N(t) \xrightarrow[N\to\infty]{a.s} 0 \qquad \text{uniformly in } t \in [0,T].$$

Hence

$$\left(\underline{H}^{(N)}(t), 0 \le t \le T\right) \xrightarrow[N\to\infty]{a.s} (\sigma \max_{0\le u\le v\le t} \{\underline{B}(v) - \underline{B}(u)\}\,;\, 0 \le t \le T). \tag{1.4.6}$$

We denote $\xi(t) = \max_{0\le u\le v\le t} \{B(v) - B(u)\} = \max_{0\le v\le t} \{B(v) - I(v)\}$ where $I(v) = \min_{0\le u\le v} B(u)$.
Recall that Paul Lévy's theorem (1948, [RY91], chap. II, thm 2.3) :

$$(B(v) - I(v), v \ge 0) \overset{(d)}{=} (|B_v|, v \ge 0)\,.$$

This ends the proof of Theorem 7. □

1.4.2 Proof of Proposition 11

This proof is similar to the previous one (see section 1.4.1) Let $H^{(N)}$ be the piecewise linear function defined by (1.4.2). The equation (1.4.3) has to be replaced by

$$\frac{1}{\sqrt{N}} S_k = \sigma_N B^{(N)}\left(\frac{k}{N}\right) + \frac{k a_N}{\sqrt{N}} = \sigma_N B^{(N)}\left(\frac{k}{N}\right) + \frac{k}{N}(\sqrt{N} a_N), \tag{1.4.7}$$

where $a_N = \sqrt{N}\mathbb{E}(X_1)$ and $\sigma_N = Var(X_1)$.
Suppose $t > 0$. Then :

$$\frac{H_{[N\,t]}}{\sqrt{N}} = \max_{0\le u\le v\le \frac{[N\,t]}{N}} \left\{ \sigma_N B^{(N)}(v) + v\,(\sqrt{N} a_N) - \sigma_N B^{(N)}(u) - u\,(\sqrt{N} a_N) \right\} \tag{1.4.8}$$

But $\sqrt{N}\,a_N$ (resp σ_N) tends to δ (resp. σ^2), the convergence follows easily.

1.4.3 Proof of Proposition 12

1) Let $\phi^{(\gamma)}(a)$ be equal to $e^{-\gamma a}\mathbb{P}(\xi_\gamma \ge a)$.
In a first step we establish the following stochastic representation for $\phi^{(\gamma)}$:

$$\phi^{(\gamma)}(a) = \mathbb{E}\left[\mathbb{1}_{\{T_a^* < 1\}} \exp\left(-\gamma\, L^0_{T_a^*} - \frac{\gamma^2}{2}\, T_a^*\right) \right], \tag{1.4.9}$$

Let f be a Borel bounded function, we have :

$$\mathbb{E}\left[f\left(\xi^{(\gamma)}\right)\right] = \mathbb{E}\left[f\left(\max_{0\le u\le 1}\left\{B_u + \gamma\, u - \min_{0\le s\le u}\left(B_s + \gamma\, s\right)\right\}\right)\right].$$

Let us apply Girsanov's theorem ([RY91], chap. VIII), we get

$$\mathbb{E}\left[f\left(\xi^{(\gamma)}\right)\right] = \mathbb{E}\left[f\left(\max_{0\le u\le 1}\left(B_u - \min_{0\le s\le u} B_s\right)\right)\exp\left\{\gamma B_1 - \frac{\gamma^2}{2}\right\}\right]. \tag{1.4.10}$$

But Levy's theorem ([RY91], chap.II) gives

$$\left(B_t - \min_{0\le s\le t} B_s, -\min_{0\le s\le t} B_s; t\ge 0\right) \overset{(d)}{=} \left(|B_t|, L_t^0; t\ge 0\right).$$

Then

$$\mathbb{E}\left[f\left(\xi^{(\gamma)}\right)\right] = \mathbb{E}\left[f\left(\max_{0\le s\le 1}|B_s|\right)\exp\left\{\gamma\left(|B_1| - L_1^0\right) - \frac{\gamma^2}{2}\right\}\right]. \tag{1.4.11}$$

Let $(M_t, t\ge 0)$ be the process :

$$M_t = \exp\left\{\gamma\left(|B_t| - L_t^0\right) - \frac{\gamma^2}{2}t\right\}; \quad t\ge 0.$$

M is an exponential martingale since $(|B_t| - L_t^0;\ t\ge 0)$ is a Brownian motion.
We restrict ourself to $f = 1_{]a,+\infty[}$, equation (1.4.11) reduces to :

$$\mathbb{P}\left(\xi_\gamma > a\right) = \mathbb{E}\left[\mathbb{1}_{\left\{B_1^*>a\right\}}\exp\left\{\gamma\left(|B_1| - L_1^0\right) - \frac{\gamma^2}{2}\right\}\right] = \mathbb{E}\left[1_{\left\{B_1^*>a\right\}}M_1\right].$$

We have $\{B_1^* > a\} = \{T_a^* < 1\}$ (recall that $B_1^* = \max_{u\le 1}|B_u|$ and $T_a^* = \inf\{t\ge 0, |B_t| > a\}$).
Let us introduce $U = T_a^* \wedge 1$. U is a bounded stopping time and $\{T_a^* < 1\} = \{U < 1\}$. Then $\{T_a^* < 1\} \in \mathcal{F}_U$, so that we may apply the stopping time theorem :

$$\begin{aligned}\mathbb{P}\left(\xi_\gamma > a\right) &= \mathbb{E}\left[\mathbb{1}_{\{T_a^*<1\}}M_1\right] = \mathbb{E}\left[\mathbb{1}_{\{T_a^*<1\}}M_U\right]\\ &= \mathbb{E}\left[\mathbb{1}_{\{T_a^*<1\}}\exp\left\{\gamma\left(a - L_{T_a^*}^0\right) - \frac{\gamma^2}{2}T_a^*\right\}\right].\end{aligned}$$

This shows (1.4.9).

2) We are now able to prove (1.3.12).
The proof is based on decomposition of Brownian path ([Val91b], prop 4).
Let us recall this decomposition :

For $a > 0$. Define

$$g = \sup \{t \leq T_a^*,\ B_t = 0\}.$$

Then

(i) $T_a^* - g$ and $(B_u,\ 0 \leq u \leq \gamma)$ are independent,

(ii) $T_a^* - g \overset{(d)}{=} T_a$,

(iii) conditionally to $L^0_{T_a^*} = t$, $(B_u,\ 0 \leq u \leq g)$ is distributed as $(B_u,\ 0 \leq u \leq \tau_t)$ conditioned by $\left\{B^*_{\tau_t} < a\right\}$.

We decompose T_a^* as the sum of g and $T_a^* - \gamma$, (1.3.12) is a straightforward consequence of (1.4.9). □

1.4.4 Second proof of Theorem 10

For simplicity $F_y^{(\lambda)}$ will be noted F_y in this section.
Let us start with a preliminary result.

Lemma 19 *Let ψ be the function :*

$$\psi(v) = \int_{\mathbb{R}_+} e^{-\gamma y} F_y\left(\frac{v}{1+v}, 0\right) \mathrm{d}y; \quad v > 0. \tag{1.4.12}$$

Then

$$\psi(v) = \frac{2}{\sqrt{2\pi}}\sqrt{\frac{v}{v+1}} + \psi_1(v), \quad |\psi_1(v)| \leq C\frac{v}{1+v}.$$

Proof : Since $F_t(x,0) = \mathbb{E}\left(\mathbb{1}_{\{0 \leq \tau_t \leq x,\}} e^{-\lambda^2 \tau_t/2}\right)$ and the density of τ_t is well known (see for example [BS96]),

$$\mathbb{P}\left(\tau_t \in \mathrm{d}z\right) = \frac{t}{\sqrt{2\pi\, z^3}} \exp\left(-\frac{t^2}{2z}\right) \mathbb{1}_{\{z \geq 0\}} \mathrm{d}z.$$

Then $F_t(x,0)$ may be written as :

$$F_t(x,0) = \frac{t}{\sqrt{2\pi}} \int_0^x \exp\left(-\frac{\lambda^2 z}{2} - \frac{t^2}{2z}\right) \frac{\mathrm{d}z}{z^{3/2}}.$$

Consequently ,

$$\psi(v) = \frac{1}{\sqrt{2\pi}} \int_0^{v/1+v} \frac{1}{z^{1/2}} e^{-\lambda^2 z/2} \int_0^{+\infty} u e^{-(u^2/2 + \gamma u\sqrt{z})} \mathrm{d}u\, \mathrm{d}z. \tag{1.4.13}$$

We have :

$$e^{-x} = 1 + \rho(x)$$

where $|\rho(x)| \leq C|x|e^{|x|}$.

In particular $e^{-\gamma u\sqrt{z}} = 1 + \rho(\gamma u\sqrt{z})$, $\psi = \psi_2 + \psi_3$ where

$$\psi_2(v) = \frac{1}{\sqrt{2\pi}} \int_0^{v/1+v} \frac{1}{z^{1/2}} e^{-\lambda^2 z/2} \left(\int_0^{+\infty} u e^{-u^2/2} \mathrm{d}u \right) \mathrm{d}z, \tag{1.4.14}$$

$$\psi_3(v) = \frac{1}{\sqrt{2\pi}} \int_0^{v/1+v} \frac{1}{z^{1/2}} e^{-\lambda^2 z/2} \left(\int_0^{+\infty} u e^{-u^2/2} \, \rho(\gamma u\sqrt{z}) \mathrm{d}u \right) \mathrm{d}z. \tag{1.4.15}$$

Clearly

$$\psi_2(v) = \frac{1}{\sqrt{2\pi}} \int_0^{v/1+v} \frac{1}{z^{1/2}} e^{-\lambda^2 z/2} \mathrm{d}z.$$

But $0 < 1 - e^{-\lambda^2 z/2} < \lambda^2 z/2$ for any $z \geq 0$, consequently

$$\psi_2(v) = \frac{2}{\sqrt{2\pi}} \sqrt{\frac{v}{v+1}} + \hat{\psi}_2(v); \quad |\hat{\psi}_2(v)| \leq C \left(\frac{v}{1+v} \right)^{3/2} \leq C \left(\frac{v}{1+v} \right).$$

But

$$|\rho(\gamma u\sqrt{z})| \leq C|\delta| u\sqrt{z} e^{|\delta| u\sqrt{z}} \leq C|\delta| \left(u e^{|\delta| u} \right) \sqrt{z}.$$

By the same way :

$$|\psi_3(v)| \leq C \left(\frac{v}{1+v} \right).$$

□

Proof of Theorem 10 : Let us recall the expression of $\phi_\lambda^{(\gamma)}$ given in equation (1.3.14).

$$\phi_\lambda^{(\gamma)}(a) = \frac{1}{a} \int_{[0,+\infty[^2} \mathbb{1}_{\{u \leq 1\}} \exp\left\{ -\gamma y - \lambda^2 u/2 \right\} \mu_a(u) \, F_y^{(\lambda)}(1-u, 1/a) \, \mathrm{d}u \, \mathrm{d}y$$

1. Let us first prove that $\phi_\lambda^{(\gamma)}(a) \underset{a\to\infty}{\sim} \rho_1(a)$, where

$$\rho_1(a) = \frac{1}{a} \int_{\mathbb{R}_+^2} \mathbb{1}_{\{u \leq 1\}} \exp\left\{ -\gamma y - \frac{\lambda^2}{2} u \right\} \mu_a(u) \, F_y^{(\lambda)}(1-u, 0) \mathrm{d}u \, \mathrm{d}y. \tag{1.4.16}$$

   Recall that

$$F_y^{(\lambda)}(1-u, 1/a) = \mathbb{E}\left[\mathbb{1}_{\left\{ 0 \leq \tau_y \leq 1-u, \, 0 \leq B^*_{\tau_y} \leq a \right\}} e^{-\lambda^2 \tau_y/2} \right],$$

   so that $\lim_{a\to\infty} F_y^{(\lambda)}(1-u, 1/a) = \mathbb{E}\left[\mathbb{1}_{\{0 \leq \tau_y \leq 1-u\}} e^{-\lambda^2 \tau_y/2} \right] = F_y^{(\lambda)}(1-u, 0)$. Since the convergence is uniform in u, taking the limit over a gives (1.4.16).

2. In this step we prove that $\rho_1(a) \underset{a\to\infty}{\sim} \rho_2(a)$, with

$$\rho_2(a) = \frac{2a^2}{\sqrt{2\pi}} \int_{\mathbb{R}_+^2} \mathbb{1}_{\{u\le 1\}} e^{-\gamma\, y - \lambda^2 u/2 - a^2/2u}\, F_y^{(\lambda)}(1-u,0) \frac{\mathrm{d}u}{u^{5/2}}\, \mathrm{d}y. \tag{1.4.17}$$

We use the explicit form of μ_a given by equation (1.3.17), the scaling property, and (1.3.16) then

$$\begin{aligned} \mu_a(u) &= \frac{1}{a^2} \frac{a^3}{\sqrt{2\pi} u^{3/2}} \sum_{k\in\mathbb{Z}} \left(-1 + a^2 \frac{(1+2k)^2}{u}\right) \exp\left\{-a^2 \frac{(1+2k)^2}{2u}\right\} \\ &= \frac{a}{\sqrt{2\pi}} \frac{R(u,a)}{u^{3/2}}, \end{aligned}$$

with

$$R(u,a) = \sum_{k\in\mathbb{Z}} \left(-1 + a^2 \frac{(1+2k)^2}{u}\right) \exp\left\{-a^2 \frac{(1+2k)^2}{2u}\right\}. \tag{1.4.18}$$

We split $R(u,a)$ in two parts :

$$\begin{aligned} R(u,a) &= 2\left(\frac{a^2}{u} - 1\right) e^{-a^2/2u} + \frac{a^2}{u} e^{-a^2/2u} \left(\sum_{k\in\mathbb{Z}-\{-1,0\}} \beta_k(u,a)\right), \\ &= \frac{2a^2}{u} e^{-a^2/2u} + \frac{a^2}{u} e^{-a^2/2u} \left(-\frac{2u}{a^2} + \sum_{k\in\mathbb{Z}-\{-1,0\}} \beta_k(u,a)\right), \end{aligned}$$

with

$$\beta_k(u,a) = \left(-\frac{u}{a^2} + (1+2k)^2\right) \exp\left\{-\frac{a^2}{2u}\left((1+2k)^2 - 1\right)\right\}.$$

We prove that the sum, k running over $\mathbb{Z}-\{-1,0\}$ goes to 0, as $a\to\infty$. If $a \ge 1$ and $u \le 1$, we have :

$$\mid -\frac{u}{a^2} + (1+2k)^2 \mid \le (1+2k)^2 + 1 \le Ck^2,$$

$$\exp\left\{-\frac{a^2}{2u}\left((1+2k)^2 - 1\right)\right\} \le \exp\{-2k(k+1)\}.$$

This yields

$$\mid \beta_k(u,a) \mid \le Ck^2 e^{-2k(k+1)}$$

The dominated convergence theorem implies that :

$$\lim_{a\to+\infty} \sum_{k\in\mathbb{Z}-\{-1,0\}} \beta_k(u,a) = 0,$$

uniformly in u.

Furthermore $\lim_{a\to\infty} \frac{u}{a^2} = 0$ uniformly with respect to $u \in [0,1]$, then :

$$R(u,a) \underset{a\to\infty}{\sim} \frac{2a^2}{u} e^{-a^2/2u}.$$

3. Finally we check that $\rho_2(a) \underset{a\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}} e^{-\lambda^2/2}\, \frac{1}{a} e^{-a^2/2}$.

We have

$$\rho_2(a) = \frac{2a^2}{\sqrt{2\pi}} \int_0^1 \frac{1}{u^{5/2}} \exp\left\{-\frac{1}{2}\left(\frac{a^2}{u} + \lambda^2 u\right)\right\} \left(\int_{\mathbb{R}_+} e^{-\gamma y}\, F_y^{(\lambda)}(1-u,0)\mathrm{d}y\right) \mathrm{d}u.$$

We set $u = \dfrac{1}{1+v}$, we obtain :

$$\rho_2(a) = \frac{2a^2}{\sqrt{2\pi}} e^{-a^2/2} \int_0^{+\infty} e^{-a^2 v/2} e^{-\lambda^2/2(1+v)} \sqrt{1+v}\, \psi(v)\mathrm{d}v, \quad (1.4.19)$$

Let us set $u = a^2 v/2$ in equation (1.4.19), then

$$\rho_2(a) = \frac{4}{\sqrt{2\pi}} e^{-a^2/2} \int_0^{+\infty} e^{-u} e^{-\lambda^2 a^2/2(a^2+2u)} \sqrt{1 + \frac{2u}{a^2}}\, \psi\left(\frac{2u}{a^2}\right) \mathrm{d}u.$$

Lemma 19 implies that :

$$\rho_2(a) = \frac{4}{\sqrt{2\pi}} e^{-a^2/2} \left[\frac{2\sqrt{2}}{\sqrt{2\pi}} \frac{1}{a} \int_0^{+\infty} e^{-u} e^{-\lambda^2 a^2/2(a^2+2u)} \sqrt{u}\mathrm{d}u + \rho_3(a)\right], \quad (1.4.20)$$

where

$$\rho_3(a) = \int_0^{+\infty} e^{-\lambda^2 a^2/2(a^2+u^2)} e^{-u} \sqrt{1 + 2u/a^2}\, \psi_1\left(\frac{2u}{a^2}\right) \mathrm{d}u.$$

The integral on the right-hand side of (1.4.20) converges as a goes to infinity to

$$e^{-\lambda^2/2} \int_0^{+\infty} e^{-u} \sqrt{u}\mathrm{d}u = e^{-\lambda^2/2} \frac{\sqrt{\pi}}{2}.$$

We claim that $|\rho_3(u)|$ is upper bounded by C/a^2, $a \to +\infty$.

Using the upper bound for ψ_1, we obtain :

$$|\rho_3(u)| \le C \left(\int_0^{+\infty} u e^{-u} \mathrm{d}u\right) \frac{1}{a^2}.$$

Finally

$$\rho_2(a) \underset{a\to\infty}{\sim} \frac{2\sqrt{2}}{\sqrt{\pi}} e^{-\lambda^2/2} \frac{1}{a} e^{-a^2/2}.$$

As $\mathbb{P}(\xi_\gamma > a) = e^{\gamma a}\, \phi_\gamma^{(\lambda)}(a)$, we have proved relation (1.3.6).

□

1.4.5 Proof of Theorem 15

We divide the proof into two steps.
1) Let $F_t^{(\lambda)}$ be the function defined by (1.3.15) :

$$F_t^{(\lambda)}(x,b) = \mathbb{E}\left(\mathbb{1}_{\left\{0\leq\tau_t\leq x,\ 0\leq B^*_{\tau_t}\leq 1/b\right\}} e^{-\lambda^2\tau_t/2}\right), \quad x\geq 0, b\geq 0, \tag{1.4.21}$$

Here λ and t are fixed. We have :

$$F_t^{(\lambda)}(x,b) = \mathbb{E}\left(\mathbb{1}_{\{0\leq\tau_t\leq x\}} e^{-\lambda^2\tau_t/2}\right) - \mathbb{E}\left(\mathbb{1}_{\left\{0\leq\tau_t\leq x,\ B^*_{\tau_t}>1/b\right\}} e^{-\lambda^2\tau_t/2}\right). \tag{1.4.22}$$

Let $B^*_{\tau_t} = u$. Let us define $\gamma = \inf\{s\leq\tau_t,\ |B_s| = u\}$, $g = \sup\{s\leq\gamma,\ B_s = 0\}$, $d = \inf\{s\geq\gamma,\ B_s = 0\}$. Vallois [Val91a] proved that conditionally to $B^*_{\tau_t} = u$,

- $g+(\tau_t - d)$,$(\gamma - g)$ et $(d-\gamma)$ are three independent random variables.
- $g + (\tau_t - d)$ is distributed as the first time when the local time of Brownian motion conditioned to stay in $[-u,u]$, reaches t.
- $\gamma - g$ and $d-\gamma$ are distributed as the first time when a Bessel process of dimension 3, started at 0, reaches u. So $(\gamma-g)+(d-\gamma)$ have same law as $T_u + \bar{T}_u$ where $\bar{T}_u$ is an independent copy of T_u.

Since $\tau_t = g + (\tau_t - d) + (\gamma - g) + (d-\gamma)$ and $\mathbb{P}\left(B^*_{\tau_t} < u\right) = e^{-t/u}$ (cf. Remark 14), we get :,

$$\begin{aligned} F_t^{(\lambda)}(x,b) &= F_t^{(\lambda)}(x,0) \\ &\quad -t\int_{1/b}^{+\infty}\frac{e^{-t/u}}{u^2}\int_0^{+\infty}\mathbb{E}\left[1_{\{\tau_t+y<x\}}e^{-(\lambda^2\tau_t)/2}|B^*_{\tau_t}<u\right] \\ &\qquad\qquad \times e^{-\lambda^2 y/2}\mu_u^{(2)}(y)\mathrm{d}y\mathrm{d}u, \\ &= F_t^{(\lambda)}(x,0) \\ &\quad -t\int_{1/b}^{+\infty}\frac{\mathrm{d}u}{u^2}\left(\int_0^{+\infty}\mathbb{E}\left[1_{\left\{\tau_t+y<x, B^*_{\tau_t}<u\right\}}e^{-\lambda^2\tau_t/2}\right]e^{-\lambda^2 y/2}\mu_u^{(2)}(y)\mathrm{d}y\right) \end{aligned}$$

We set $v = 1/u$, (1.3.20) follows immediately since we have already established (1.3.15) in the proof of Lemma 19.
2) Let K be a positive number and E_K the set of Borel functions ψ defined on $\mathbb{R}_+\times[0,K]$ such that

$$\sup_{x\geq 0, y\leq K}|\psi(x,y)| < +\infty.$$

E_K is equipped with the uniform norm.

Let ψ be in E_K, $x \geq 0$ and $a \leq K$. Then

$$\begin{aligned} |A^{(\lambda)}\psi\,(x,a)| &\leq \int_0^a \mathrm{d}u \left(\int_0^x \mu_{1/u}^{(2)}(y)\, e^{-\lambda^2 y/2}\, |\psi(x-y,u)| \mathrm{d}y \right) \\ &\leq \max_{s\geq 0,\, 0\leq u\leq a} |\psi(s,u)| \int_0^a \mathrm{d}u \left(\int_0^x \mu_{1/u}^{(2)}(y)\, \mathrm{d}y \right), \end{aligned}$$

$\mu_{1/u}^{(2)}$ being a density function :

$$|A^{(\lambda)}\psi\,(x,a)| \leq K \max_{s\geq 0,\, 0\leq u\leq K} |\psi(s,u)|.$$

$A^{(\lambda)}$ is thus a continuous linear operator from E_K to E_K.

Clearly $(x,a) \mapsto F_t^{(\lambda)}(x,0)$ belongs to E_K, because

$$0 \leq F_t^{(\lambda)}(x,0) \leq 1. \tag{1.4.23}$$

Let us consider the series

$$\Lambda_t(x,a) = \sum_{k=0}^{+\infty} (-1)^k\, t^k \alpha_t^{(k)}(x,a), \tag{1.4.24}$$

with

$$\alpha_t^{(0)}(x,a) = F_t^{(\lambda)}(x,0)$$

and

$$\alpha_t^{(k+1)}(x,a) = \left(A^{(\lambda)} \alpha_t^{(k)} \right)(x,a).$$

In order to establish the convergence in E_K, we first prove that

$$\max_{x\geq 0, y\leq a} |\alpha_t^{(k)}(x,y)| \leq \frac{a^k}{k!} \max_{x\geq 0, y\leq a} |\alpha_t^{(0)}(x,y)| \leq \frac{a^k}{k!}. \tag{1.4.25}$$

We check (1.4.25) by induction on n.
If $n = 0$, obviously (1.4.25) holds. We suppose that (1.4.25) is verified for n and we prove that (1.4.25) is still true, having replaced n by $n+1$.
Let $x \geq 0$, $0 \leq y \leq a$, using again the fact that $\mu_{1/u}^{(2)}$ is a density function, we obtain

$$\begin{aligned} |A^{(\lambda)}\psi(x,y)| &\leq \int_{[0,+\infty[^2} \mathbb{1}_{\{u\leq y\}}\, \mu_{1/u}^{(2)}(v) \max_{x\geq 0} |\psi(x,u)|\, \mathrm{d}v\, \mathrm{d}u, \\ &\leq \int_0^y \max_{x\geq 0} |\psi(x,u)|\, \mathrm{d}u \leq \int_0^a \left(\max_{x\geq 0,\, u_1\leq u} |\psi(x,u_1)| \right) \mathrm{d}u. \end{aligned}$$

Therefore

$$\max_{x\geq 0, y\leq a} |A^{(\lambda)}\psi(x,y)| \leq \int_0^a \left(\max_{x\geq 0,\, u_1\leq u} |\psi(x,u_1)| \right) \mathrm{d}u. \tag{1.4.26}$$

Consequently (1.4.25) implies

$$\max_{x\geq 0,\, u\leq a} |\alpha^{(n+1)}(x,u)| \leq \frac{1}{n!}\int_0^a u^n \,\mathrm{d}u = \frac{a^{n+1}}{(n+1)!}.$$

Consequently the series in (1.4.24) converge in E_K, $A^{(\lambda)}$ is a continuous operator, then

$$F_t^{(\lambda)}(x,a) = \sum_{k=0}^{+\infty} (-1)^k\, t^k \alpha_t^{(k)}(x,a), \quad (x,a)\in\mathbb{R}_+^2.$$

3) Recall the expression of $\phi_\lambda^{(\gamma)}$ in terms of $F_t^{(\lambda)}$.

$$\phi_\lambda^{(\gamma)}(a) = \frac{1}{a}\int_{[0,+\infty[^2} \mathbb{1}_{\{u\leq 1\}}\, e^{-\gamma t-\lambda^2 u/2}\, F_t^{(\lambda)}\,(1-u,1/a)\;\mu_a(u)\,\mathrm{d}u\,\mathrm{d}t$$

We are allowed to interchange the sum with respect to k and the double integral if :

$$\sum_{k\geq 1}\beta_k < +\infty$$

with

$$\beta_k = \frac{1}{a}\int_{[0,+\infty[^2} \mathbb{1}_{\{u\leq 1\}} e^{-\gamma t-\lambda^2 u/2}\mu_a(u) t^k \alpha_t^{(k)}(1-u,1/a)\mathrm{d}u\,\mathrm{d}t.$$

It is well known that $\tau_t \overset{(d)}{=} t^2/B_1^2$, then if $x\leq 1$, $t>0$,

$$0\leq \alpha_t^{(0)}(x,a) \leq \mathbb{P}(\tau_t\leq x) \leq \mathbb{P}(\tau_t\leq 1) = 2\mathbb{P}(B_1>t) \leq \frac{2}{\sqrt{2\pi t}} e^{-t^2/2}.$$

Obviously (1.4.26) can be modified as follows :

$$\max_{0\leq x\leq 1, y\leq a} |A^{(\lambda)}\psi(x,y)| \leq \int_0^a \left(\max_{0\leq x\leq 1,\, u_1\leq u} |\psi(x,u_1)|\right)\mathrm{d}u.$$

Reasoning by induction, we obtain :

$$\max_{0\leq x\leq 1, u\leq a} |\alpha_t^{(k)}(x,u)| \leq \frac{2}{\sqrt{2\pi t}} e^{-t^2/2}\frac{a^k}{k!}.$$

Consequently

$$\beta_k \leq \frac{2}{a\sqrt{2\pi}}\int_0^{+\infty} e^{-\gamma t-t^2/2}\left(\frac{t}{a}\right)^k \frac{1}{k!}\frac{\mathrm{d}t}{\sqrt{t}}.$$

$$\sum_{k\geq 1}\beta_k \leq \frac{2}{a\sqrt{2\pi}}\int_0^{+\infty} e^{-\gamma t-t^2/2}\left(e^{t/a}-1\right)\frac{\mathrm{d}t}{\sqrt{t}} < +\infty.$$

This implies the identity (1.3.27). □

1.4.6 Proof of formula (1.2.7).

Recall that $(B_t, t \geq 0)$ is a standard Brownian motion, and $B_1^* = \max_{0\leq s\leq 1} |B_s|$. The cumulative function of B_1^* is known (cf [BS96], p.146) :

$$\mathbb{P}\left(B_1^* < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} \sum_{k\in\mathbb{Z}} \left(e^{-\frac{(y+4kx)^2}{2}} - e^{-\frac{(y+2x+4kx)^2}{2}} \right) \mathrm{d}y. \tag{1.4.27}$$

Jacobi's theta function identity ([Bel61]) gives us :

$$\frac{1}{\sqrt{\pi t}} \sum_{k\in\mathbb{Z}} e^{-\frac{(v+k)^2}{t}} = \sum_{k\in\mathbb{Z}} \cos\left(2k\pi v\right) e^{-k^2\pi^2 t}, \quad v \in \mathbb{R},\ t > 0. \tag{1.4.28}$$

Setting $v = y/4x$ and $t = 1/8x^2$,(1.4.28) becomes :

$$\frac{4x}{\sqrt{2\pi}} \sum_{k\in\mathbb{Z}} e^{-\frac{(4kx+y)^2}{2}} = \sum_{k\in\mathbb{Z}} \cos\left(k\pi y/2x\right) e^{-\frac{k^2\pi^2}{8x^2}}. \tag{1.4.29}$$

Then

$$\frac{1}{\sqrt{2\pi}} \sum_{k\in\mathbb{Z}} e^{-\frac{(4kx+y)^2}{2}} = \frac{1}{4x} \sum_{k\in\mathbb{Z}} \cos\left(k\pi y/2x\right) e^{-\frac{k^2\pi^2}{8x^2}}. \tag{1.4.30}$$

Similarly, setting $v = (y + 2x)/4x$ and $t = 1/8x^2$ in (1.4.28), we obtain :

$$\frac{1}{\sqrt{2\pi}} \sum_{k\in\mathbb{Z}} e^{-\frac{(4kx+2x+y)^2}{2}} = \frac{1}{4x} \sum_{k\in\mathbb{Z}} (-1)^k \cos\left(k\pi y/2x\right) e^{-\frac{k^2\pi^2}{8x^2}}. \tag{1.4.31}$$

Integrating in y, we obtain the cumulative distribution for B_1^* :

$$\mathbb{P}(B_1^* < x) = \frac{2}{\pi} \sum_{k\in\mathbb{Z}} \frac{(-1)^k}{2k+1} e^{-\frac{(2k+1)^2\pi^2}{8x^2}}. \tag{1.4.32}$$

1.4.7 Proof of formula (1.3.17).

Let us denote by $(R_x(s), s \geq 0)$ a Bessel process of dimension 3 starting at x and $T_a^{(x)}$ the first time where $(R_x(s))_{s\geq 0}$ reaches a ($T_a^{(x)} = \inf\{t \geq 0; R_x(t) = a\}$).
We claim that $T_a^{(0)}$ admits μ_a as a density function, where

$$\mu_a(t) = \frac{1}{a^2} \mu_1\left(\frac{t}{a^2}\right),$$

$$\mu_1(t) = \frac{1}{\sqrt{2\pi} t^{3/2}} \sum_{k\in\mathbb{Z}} \left(-1 + \frac{(1+2k)^2}{t}\right) \exp - \frac{(1+2k)^2}{2t}. \tag{1.4.33}$$

In [BS96], (page 339, 2.02) we find the density function of $T_a^{(x)}$, for $0 < x < a$:

$$P\left(T_a^{(x)} \in dt\right) = \frac{a}{x}\Psi_x^{(a)}(t)\mathbb{1}_{\{t\geq 0\}}\,dt = \varphi_x^{(a)}(t)\mathbb{1}_{\{t\geq 0\}}\,dt \tag{1.4.34}$$

where

$$\Psi_x^{(a)}(t) = \frac{1}{\sqrt{2\pi t^3}}\sum_{k\in\mathbb{Z}}(a - x + 2ka)\exp -\frac{(a-x+2ka)^2}{2t}. \tag{1.4.35}$$

Let us prove that $\Psi_0^{(a)}(t) = 0$.
For all $t > 0$, we have :

$$\begin{aligned}\Psi_0^{(a)}(t) =& \frac{a}{\sqrt{2\pi t^3}}\sum_{k\in\mathbb{Z}}(1+2k)e^{-\frac{(1+2k)^2\,a^2}{2t}}\\ =& \frac{a}{\sqrt{2\pi t^3}}\left\{\sum_{k=0}^{+\infty}(1+2k)e^{-\frac{(1+2k)^2\,a^2}{2t}} + \sum_{k=0}^{+\infty}(1+2(-k-1))e^{-\frac{(1+2(-k-1)k)^2\,a^2}{2t}}\right\},\\ =& 0\end{aligned}$$

Then

$$\mu_a(t) = \lim_{x\to 0}\varphi_x^{(a)}(t) = \lim_{x\to 0}\frac{a}{x}\left(\Psi_x^{(a)}(t) - \Psi_0^{(a)}(t)\right) \tag{1.4.36}$$

Differentiating term by term, we obtain :

$$\mu_a(t) = \frac{a}{\sqrt{2\pi}\,t^{3/2}}\sum_{k\in\mathbb{Z}}\left(-1 + \frac{a^2(1+2k)^2}{t}\right)\exp -\frac{a^2(1+2k)^2}{2t} \tag{1.4.37}$$

□

1.4.8 Proof of (1.3.18).

We make use of Poisson formula ([Fel66], chap. XIX, p.620).

$$\sum_{k\in\mathbb{Z}}\varphi(a+2kb) = \frac{\pi}{b}\sum_{k\in\mathbb{Z}}f\left(\frac{k\pi}{b}\right)\exp\left(\frac{ik\pi a}{b}\right). \tag{1.4.38}$$

with

$$\varphi(\alpha) = \int_{\mathbb{R}} e^{i\alpha x}f(x)\mathrm{d}x.$$

We choose

$$f(x) = \sqrt{\frac{t}{2\pi}}\exp -\frac{t}{2}\left(x - \frac{\pi}{2}\right)^2.$$

f is the density function of $\frac{\pi}{2} + \frac{B_1}{\sqrt{t}}$, t being a fixed number, then :

$$\varphi(\alpha) = e^{i\alpha\pi/2} e^{-\alpha^2/2t}.$$

We set $a = 0$ and $b = 1$ in (1.4.38), we obtain :

$$\sum_{k\in\mathbb{Z}} (-1)^k e^{-2k^2/t} = \sqrt{\frac{\pi t}{2}} \sum_{k\in\mathbb{Z}} \exp -\frac{t}{8}(2k-1)^2 \pi^2.$$

We set $t = \dfrac{4}{u\pi^2}$:

$$\sum_{k\in\mathbb{Z}} (-1)^k e^{-k^2\pi^2 u/2} = \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{u}} \sum_{k\in\mathbb{Z}} \exp -\frac{(2k-1)^2}{2u}. \tag{1.4.39}$$

Differentiating in respect to u, we obtain (1.3.18). □

1.4.9 Proof of formula (1.3.23).

We keep the notations introduced in the beginning of 1.4.7.
Let us recall that $\mu_1^{(2)}$ is the density function of $Z = T_1^{(0)} + \hat{T}_1^{(0)}$, $\hat{T}_1^{(0)}$ being an independent copy of $T_1^{(0)}$.
The Laplace transform of $T_1^{(0)}$ is well known ([Ken78]) :

$$\mathbb{E}\left(e^{-\lambda T_1^{(0)}}\right) = \frac{\sqrt{2\lambda}}{sh\sqrt{2\lambda}}.$$

So that

$$\mathbb{E}\left(e^{-\lambda Z}\right) = \left(\frac{\sqrt{2\lambda}}{sh\sqrt{2\lambda}}\right)^2.$$

According to prop. 1, p.7 in [BPY01], this is equivalent to :

$$\sqrt{\frac{\pi}{2}} Z \overset{(d)}{=} Y,$$

where

$$\mathbb{P}(Y \le y) = \frac{4\pi}{y^3} \sum_{n\ge 1} n^2 e^{-\pi n^2/y^2}$$

A straightforward computation implies (1.3.23). □

Bibliography

[Da02] Daudin J.J. and Etienne M.P. . Comparisons of three approximations of the local score. *to appear*, 2002.

[Bel61] R. Bellman. *A Brief Introduction to Theta Functions.* Holt, Rinehart and Winston, 1961.

[Bil68] P. Billingsley. *Convergence of probability measures.* John Wiley and Sons, 1968. New York.

[BPY01] Philippe Biane, Jim Pitman, and Marc Yor. Probability laws related to the Jacobi theta and Riemann zeta functions, and Brownian excursions. *Bull. Amer. Math. Soc. (N.S.)*, 38(4):435–465 (electronic), 2001.

[BS96] A. N. Borodin and P. Salminen. *Handbook of Brownian motion – Facts and formulae.* Birkhauser Verlag, 1996. Basel.

[DK92] A. Dembo and S. Karlin. Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Prob*, 24:113–140, 1992. USA.

[DM99] J.J. Daudin and S. Mercier. Distribution exacte du score local d'une suite de variables indépendantes et identiquement distribuées. *C. R. Acad. Sciences*, 9:815–820, 1999. Série I, Math.

[Fel66] W. Feller. *An introduction to probability theory and its applications.* John Wiley and Sons, 1966. New York.

[Igl72] D. L. Iglehart. Extreme Values in the GI/G/1 queue. *The annals of Mathematical Statistics*, 43:627–635, 1972. USA.

[IW81] N. Ikeda and S. Watanabe. *Stochastic Differential Equations and Diffusion Processes.* North-Holland Publishing Company, 1981. Amsterdam, New-York, Oxford.

[Ken78] J. Kent. Some probabilistic properties of Bessel functions. *Annals of Probability*, 6:760–770, 1978.

[RY91] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion.* Springer Verlag, 1991. Berlin.

[Tay75] Howard M. Taylor. A stopped Brownian motion formula. *Ann. Probability*, 3:234–246, 1975.

[Val91a] P. Vallois. Sur la loi conjointe du maximum et de l'inverse du temps local du mouvement brownien. Application à un théorème de Knight. *Stochastics and Stochastics Reports*, 35:175–186, 1991.

[Val91b] P. Vallois. Une extension des théorèmes de Ray et Knight sur les temps locaux Browniens. *Probab. Theory Relat. Fields*, 88, No.4:445–482, 1991.

[Wat95] G. N. Watson. *A treatise on the theory of Bessel functions*. Cambridge University Press, Cambridge, 1995. Reprint of the second (1944) edition.

[Wid41] David Vernon Widder. *The Laplace Transform*. Princeton University Press, Princeton, N. J., 1941.

[Wil76] David Williams. On a stopped Brownian motion formula of H. M. Taylor. In *Séminaire de Probabilités, X (Première partie, Univ. Strasbourg, Strasbourg, année universitaire 1974/1975)*, pages 235–239. Lecture Notes in Math., Vol. 511. Springer, Berlin, 1976.

Chapitre 2

Comparisons of three approximations for the local score when $E(X) \simeq 0$

Comparisons of three approximations for the local score when $E(X) \simeq 0$

Jean-Jacques DAUDIN [a], Marie Pierre ETIENNE[b].

[a]Institut National Agronomique Paris-Grignon,
Département OMIP, UMR INAPG-INRA, 96021111,
16, rue C. Bernard, 75231 Paris Cedex 05, France.
E-mail : daudin@inapg.inra.fr

[b]Institut de Mathématiques Elie Cartan, Université Henri Poincaré.
BP. 239, 54506 Vandoeuvre Lès Nancy Cedex, France.
E-mail : Marie-Pierre.Etienne@iecn.u-nancy.fr

Abstract

Let $(X_i)_{i\geq 1}$ be a sequence of random variables. The local score of (X) is defined by $H_n = \max_{0\leq i\leq j\leq n}(X_i + \ldots + X_j)$. There exist several approximations to the p.d.f of H_n. The well known Dembo and Karlin's extreme value approximation (valid for $E(X) < 0$) and two new ones based on a Brownian motion approach (valid when $E(X) \simeq 0$). This paper is devoted to the comparison of them when $E(X) \simeq 0$ and attempts to give a more precise scope of validity for each one.

Keywords : Local score, Brownian motion, asymptotic behaviour.
AMS 1991 Subject classifications
60G50

2.1 Introduction

Let $(X_i)_{i\geq 1}$ be a sequence of random variables. Let

$$S_k = X_1 + \ldots + X_k, \quad S_0 = 0. \tag{2.1.1}$$

The local score of (X) is defined as followed :

$$H_n = \max_{0\leq i\leq j\leq n}(S_j - S_i) = \max_{0\leq j\leq n}\left(S_j - \min_{0\leq i\leq j} S_i\right) \tag{2.1.2}$$

The local score is used for biological sequence analysis, see [KA90]. When the X_i are i.i.d. with negative expectation, Dembo and Karlin have shown ([DK92]) :

$$\lim_{n\to+\infty} \mathbb{P}\left(H_n \leq \frac{\ln n}{\lambda} + x\right) = \exp\left(-K^* \exp\left(-\lambda x\right)\right) \tag{2.1.3}$$

where K^* and λ depend only on the probability distribution of X_1.

Actually the preceding result is true only for continuous X_1. In the lattice case the same authors have obtained two bounds and more recently Bacro, Daudin, Mercier and Robin [BDMR00] have given the following sharper bounds

$$\begin{aligned} \exp\left(-\frac{\delta}{1-R}\theta R^x\right) &\leq \liminf_{n\to\infty} P[H_n \leq \frac{\log n}{-\log(R)} + x] \\ \limsup_{n\to\infty} P[H_n &\leq \frac{\log n}{-\log(R)} + x] \leq \exp\left(-\frac{\delta}{1-R}\theta R^{x+1}\right) \end{aligned}$$

where R, δ and θ depend only on the probability distribution of X_1.

From a quite different approach, Daudin and Mercier [DM99] have obtained the exact value of $\mathbb{P}\left(H_n < x\right)$ for any n, x, without restriction on $\mathbb{E}\left(X_1\right)$. However $\mathbb{P}\left(H_n < x\right)$ is expressed through a vector P_n of length x given by :

$$P_n = P_0 \Pi^n$$

where $P_0 = (1, 0, \ldots, 0)$ of length x and π is a transition matrix $x \times x$. Then the distribution of the local score is given by :

$$\mathbb{P}\left(H_n \geq a\right) = P_n(a). \tag{2.1.4}$$

In practice, this formula is available if n and x are not too large.

Daudin, Etienne and Vallois [DEV00] have shown in the centred case :

$$\mathbb{P}\left(\frac{H_n}{\sqrt{n}} \geq a\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(\sigma B_1^* \geq a\right), \tag{2.1.5}$$

where $B_1^* = \max_{0\leq u\leq 1} |B_u|$, and $(B_u, u \geq 0)$ denotes a standard Brownian motion started at 0.

Last Daudin Etienne and Vallois have obtained another approximation only valid for the tail of the distribution, i.e. if a is large enough:

$$\mathbb{P}\left(H_n \geq a\right) \underset{a\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}} \frac{\sigma\sqrt{n}}{a} \exp - \left(\delta_n^2/(2\sigma^2)\right) \exp\left(\frac{\delta_n a}{\sigma^2\sqrt{n}} - \frac{a^2}{2\sigma^2 n}\right). \tag{2.1.6}$$

Finally we can use three approximations ((2.1.3), (2.1.5), (2.1.6)) and we also have the exact expression ((2.1.4)). In this paper we make some comparisons between the approximations (2.1.3) and (2.1.5)for means between δ_n (with $\delta_n < 0$ possibly depending on n)and 0 using (2.1.4) as reference in order to give some keys for choosing the best one. When $E(X_1) = 0$ (2.1.5) is

obviously the best choice. When $E(X_1) << 0$, (2.1.3) is better. But when $E(X_1) < 0$, and near from 0, which is a quite frequent case, nobody knows what to do. The aim of this paper is to give some indication about the phase transition point between the logarithmic and the square root behaviour of the local score. First of all we study the behaviour of (2.1.3), which is the most largely used approximation, when the mean is 0 and in the case of a slightly negative mean. The third part of this paper is devoted to the study of the behaviour of the two Brownian approximations ((2.1.5) and (2.1.6)). In the last part we try to give some practical criterion to choose the best approximation in a given context.

2.2 Limits of Karlin's approximation

At the present time Karlin's result (2.1.3) is mostly used to compute the score over long sequences : by instance the software BLAST uses it to compute the "P-value" and the "E-value". But this result is only correct for negative mean, and gives quite misleading results in the centered case or for a slightly negative mean as illustrated in the two following sections.

2.2.1 The centered case

The figure 2.1 compares the exact distribution and Karlin approximation for Bernoulli random variables, with mean 0. Using Karlin approximation in this case is clearly a bad choice.

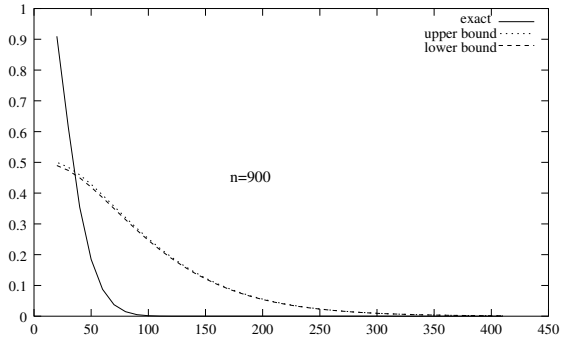

Figure 2.1: Karlin approximation,$\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 0.5$, $\mathbb{E}(X) = 0$, $n = 900$.

The figure 2.2 gives the same probabilities for the same distribution and for $n = 2500$

The last example of this section (figure 2.3) deals with another type of distribution and length $n = 900$. The p.d.f. of X is:

x_i	-3	-1	1	3
$\mathbb{P}(X = x_i)$	0.3	0.2	0.2	0.3

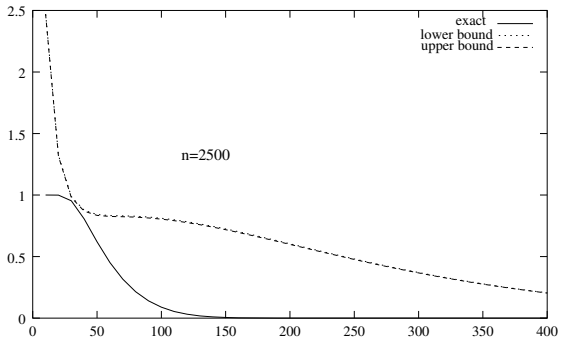

Figure 2.2: Karlin approximation,$\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(X = -1) = 0.5$, $n = 2500$.

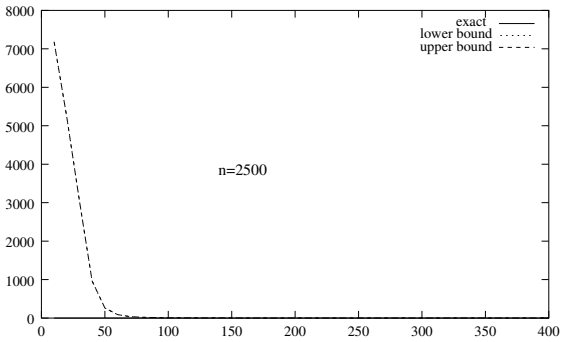

Figure 2.3: Karlin approximation,$X = -3, -1, 1$ or 3 ,$E(X) = 0$, $n = 2500$.

It appears clearly on these three examples that Karlin's approximation should not be used for a null mean for any value of n and for any p.d.f. of X.

2.2.2 A slightly negative mean

It is well known that there is a phenomena of phase transition between the logarithmic and the square root behaviour of the local score. How does (2.1.3) works when $E(X) < 0$ and is near from 0 ? In order to investigate this point we have used three negative values for $E(X)$: -0.002, 0.02, and 0.2 with a simple Bernoulli distribution. More general multinomial p.d.f. for X have been used with quite similar results and are not presented here. Note that Karlin approximation uses extreme values, so it may be bad for small a and quite good for large values of a.

Bernoulli variables with $E(X) = -0.002$.

As we have done in the previous section, it may be interesting to study the role of the length of the sequence. Does it change the global behaviour ?

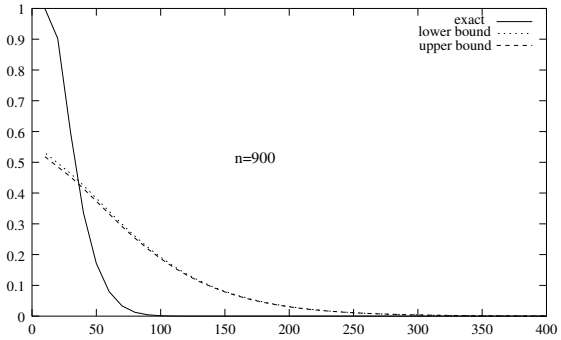

Figure 2.4: Karlin approximation,$\mathbb{P}(X = -1) = 0.501$, $\mathbb{P}(X = 1) = 0.499$, $E(X) = -0.002$, $n = 900$.

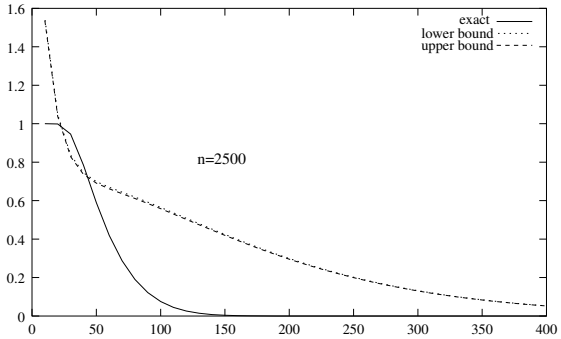

Figure 2.5: Karlin approximation,$\mathbb{P}(X = -1) = 0.501$, $\mathbb{P}(X = 1) = 0.499$, $E(X) = -0.002$, $n = 2500$.

From the two figures 2.4 and 2.5 we can see that the value of n has almost no influence on the quality of the approximation. Even if the behaviour is a bit different, the error made using Karlin's result in these cases is not acceptable.

Bernoulli variables with $E(X) = -0.02$.

The p.d.f. of X is given by

$$\mathbb{P}(X = -1) = 0.51, \quad \mathbb{P}(X = 1) = 0.49, \quad \mathbb{E}(X) = -0.02.$$

For $n = 900$, the parameter $\delta_{900} = \sqrt{n}E(X) = -0.6$, the behaviour is given in the figure 2.6.

With the same distribution for the random variables, let us study the score over longer sequence, $n = 2500$.

It appears on figure 2.6 and 2.7 that the approximation is better than in the first case, furthermore it seems to be sensible to the length of the sequence. As we have already written, this result uses extreme values, so it

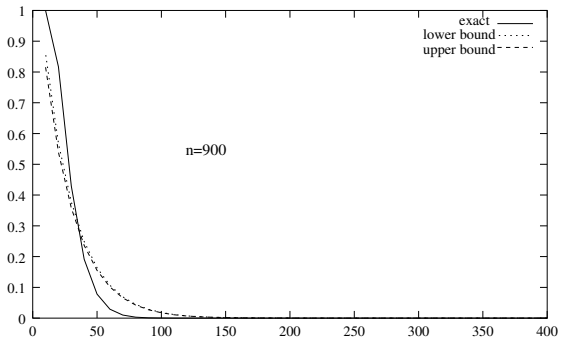

Figure 2.6: Karlin approximation,$\mathbb{P}(X = -1) = 0.51$, $\mathbb{P}(X = 1) = 0.49$, $E(X) = -0.02$, $n = 900$.

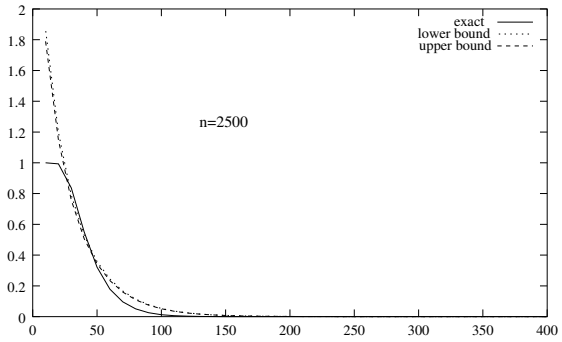

Figure 2.7: Karlin approximation,$\mathbb{P}(X = -1) = 0.51$, $\mathbb{P}(X = 1) = 0.49$, $E(X) = -0.02$, $n = 2500$.

is not surprising that it isn't so good for low value of a. If we make a zoom for the tail of the distribution (figure 2.8), we may conclude that a mean of -0.02 is not negative enough to use safely the approximation (2.1.3). It overestimates the p-value. So some rare events are not detected because of this overestimation of the distribution function.

Bernoulli variables with $E(X) = -0.2$.

In this case Karlin's approximation is correct as it can be seen in Figures 2.9, 2.10 and 2.11.

Finally we have determined a domain on which one may use safely Karlin's result. But what can we do when the mean is negative and greater than -0.2 ? The following section is devoted to the study of two approximations based on the results of Daudin, Etienne and Vallois ([DEV00]). We compare the exact distribution of the local score with the Brownian approximations for the same cases of this section and some more cases.

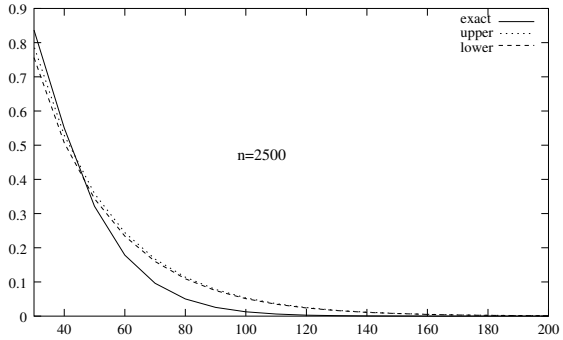

Figure 2.8: Karlin approximation,Zoom on 2.7

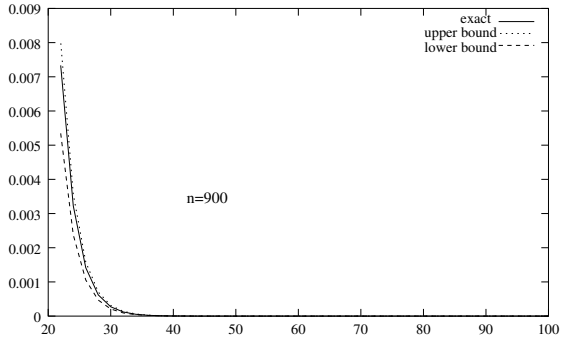

Figure 2.9: Karlin approximation, $\mathbb{P}(X = -1) = 0.6$, $\mathbb{P}(X = 1) = 0.4$, $E(X) = -0.2$, $n = 900$.

2.3 Brownian approximation

2.3.1 The centred case

Let us present our results about the Brownian approximation. Let $M = max(X)$. We use the exact p.d.f. of H_n and a lower and upper bounds:

- We approximate $\mathbb{P}(H_n > a\sqrt{n})$ with its limit $\mathbb{P}(\sigma B_1^* > a)$. Numerical results show us that $\mathbb{P}(H_n > a\sqrt{n}) \leq \mathbb{P}(\sigma B_1^* > a)$
- $\mathbb{P}(H_n > a\sqrt{n}) = P_n(a\sqrt{n})$,
- $\mathbb{P}(H_n > a\sqrt{n}) \geq \mathbb{P}(\sigma B_1^* > a + M/\sqrt{n})$,

We have not be able to prove that the lower and upper bounds are effective. However all our numerical computations agree with this statement. In the following table the first column contains n, the first line contains a and each cell is divided in three parts containing respectively in this order the upper bound, $\mathbb{P}(\sigma B_1^* > a)$, the exact value, $P_n(a\sqrt{n})$ and the lower bound

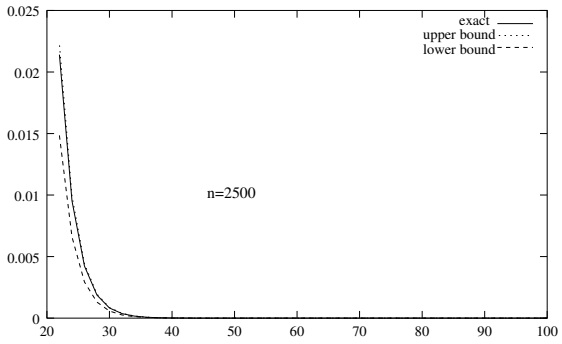

Figure 2.10: Karlin approximation,$\mathbb{P}(X = -1) = 0.6$, $\mathbb{P}(X = 1) = 0.4$, $E(X) = -0.2$, $n = 2500$.

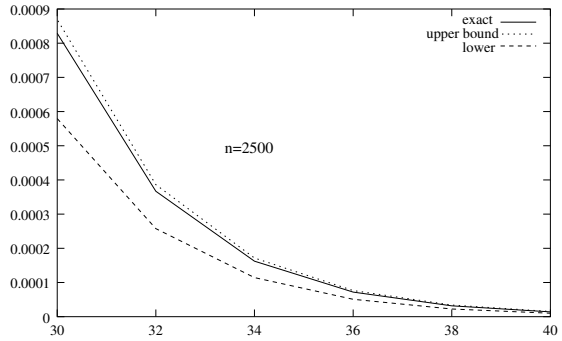

Figure 2.11: Karlin approximation,zoom, $\mathbb{P}(X = -1) = 0.6$, $\mathbb{P}(X = 1) = 0.4$, $E(X) = -0.2$, $n = 2500$.

$\mathbb{P}\left(\sigma B_1^* > a + M/\sqrt{n}\right)$.

	1	2	3	4	5
100	0 .6292	0.0910	0.0054	1.2668e-04	1.1466e-06
	0.5879	0.0812	0.0044	8.7491e-05	5.5331e-07
	0.5407	0.0715	0.0039	8.2630e-05	6.7931e-07
900	0.6292	0.0910	0.0054	1.2668e-04	1.1466e-06
	0.6145	0.0875	0.0051	1.1611e-04	1.0028e-06
	0.5990	0.0840	0.0048	1.0998e-04	9.6405e-07
10000	0.6292	0.0910	0.0054	1.2668e-04	1.1466e-06
	0.6247	0.0899	0.0053	1.2385e-04	1.1306e-06
	0.6201	0.0889	0.0052	1.2144e-04	1.1172e-06

We thus can see that the Brownian approximation works well in the centered case.

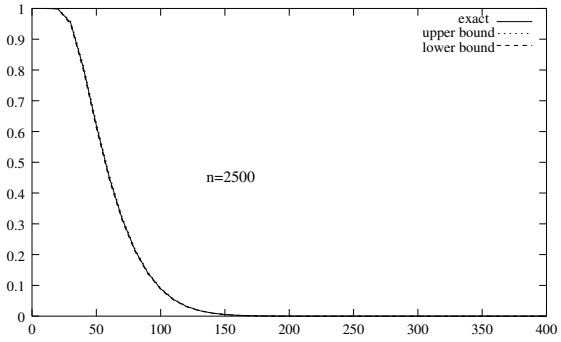

Figure 2.12: Centered Brownian approximation,$\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 0.5$, $E(X) = 0$, $n = 2500$

2.3.2 A first approximation for the non centred case

Daudin Etienne et Vallois ([DEV00]) have studied the asymptotic behaviour of the local score for any mean. Recall their result in the non null mean case :

$$\mathbb{P}\left(\frac{H_n}{\sqrt{n}} > a\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(\max_{0\le u\le 1}\left\{\sigma B(u) + \delta_n u - \min_{0\le s\le u}(\sigma B(s) + \delta_n s)\right\} > a\right). \tag{2.3.1}$$

where $\delta_n = \sqrt{n}\mathbb{E}(X_i)$, n being the length of the sequence. Although the distribution function of $\max_{0\le u\le 1}\{\sigma B(u) + \delta_n u - \min_{0\le s\le u}(\sigma B(s) + \delta_n s)\}$ is given in [DEV00], in practice it is not computable. Formula (2.1.5) may be written as

$$\mathbb{P}\left(\frac{H_n}{\sqrt{n}} \ge a\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(\max_{0\le u\le 1}\left\{\sigma B(u) - \min_{0\le s\le u}\sigma B(s)\right\} \ge a\right). \tag{2.3.2}$$

If δ_n is small, it is clear that these two formulae are very close. A natural idea is to use (2.1.5) in order to approach (2.3.1) as δ_n is near from 0. Now the natural question is : for which values of δ_n is this approximation correct ?

The aim of this section is then to determine a domain on which this approximation is valid. We will use the same approach as in the previous section.

Note that this approximation is valid for positive and negative values of $E(X)$.

A small positive mean.

The bounds used in the centered case are no more valid and it is clear on figure 2.13 that we have a small overestimation with $\delta_{900} = 0.06$ and a large one when $\delta_{900} = 0.6$ (2.14).

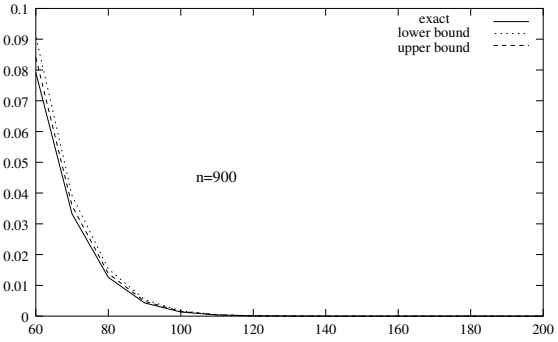

Figure 2.13: Centered Brownian approximation,$\mathbb{P}(X = 1) = 0.501$, $\mathbb{P}(X = -1) = 0.499$, $E(X) = 0.002$, $n = 900$, $\delta_{900} = 0.06$

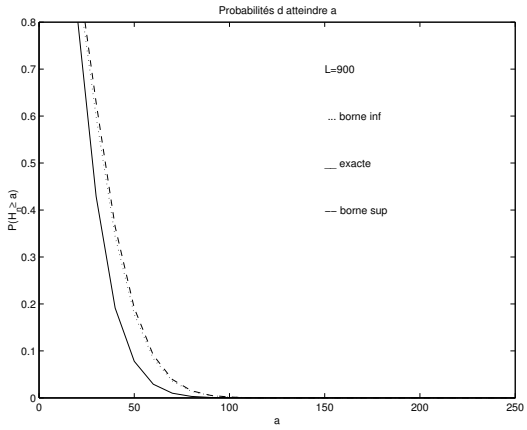

Figure 2.14: Centered Brownian approximation,$\mathbb{P}(X = 1) = 0.51$, $\mathbb{P}(X = -1) = 0.49$, $E(X) = 0.02$, $n = 900$, $\delta_{900} = 0.6$

We now give the behaviour in the same conditions as done for Karlin's approximation, in order to compare them. Like in the positive case, bounds are not still valid, so we only compute the first one, (2.1.5).

Bernoulli variables with $E(X) = -0.002$.

We can see on figure 2.15 that the Brownian approximation in this case is quite better than Karlin's one. The error made is very small, so in this case, we have to prefer (2.1.5).

When the size of the sequence is smaller ($n = 900$), Figure 2.16 shows that (2.1.5) is even better.

For Karlin's formula the value of the mean $E(X)$ is the important parameter which governs the quality of the approximation. On the opposite the crucial parameter for (2.1.5) is δ_n. The lowest δ_n, the best is the approx-

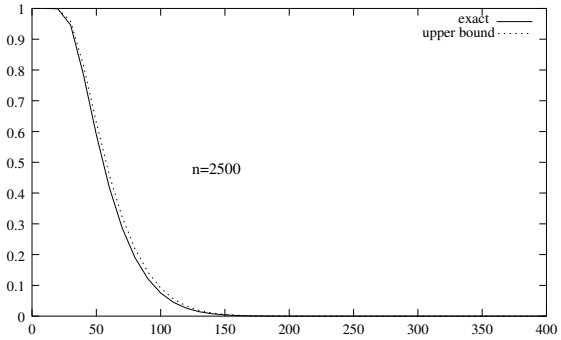

Figure 2.15: Centered Brownian approximation, $\mathbb{P}(X = -1) = 0.501$, $\mathbb{P}(X = 1) = 0.499$, $E(X) = -0.002$, $n = 2500$, $\delta_{2500} = -0.1$

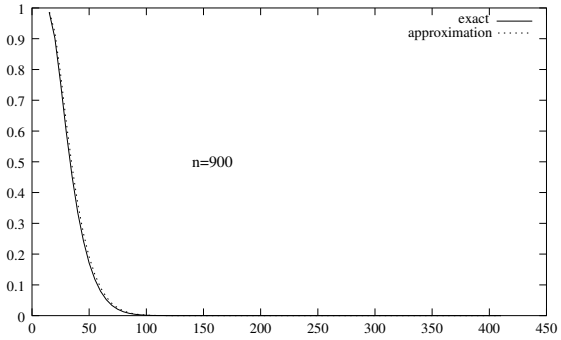

Figure 2.16: Centered Brownian approximation, $\mathbb{P}(X = -1) = 0.501$, $\mathbb{P}(X = 1) = 0.499$, $E(X) = -0.002$, $n = 900$, $\delta_{900} = -0.06$

imation, a fact which is easily explained: we approach a Brownian motion with drift δ_n by a standard Brownian motion and of course if the drift is very small, behaviours are very close.

Bernoulli variables with $E(X) = -0.02$.

In order to determine if this approximation is still valid for higher absolute values of the parameter δ_n, we have computed the case $n = 900$ with $E(X) = -0.02$, i.e $\delta_{900} = -0.6$.

Even if the two lines seem to be close from each other, a zoom shows (figure 2.18) that the error made is quite important.

2.3.3 An approximation for the tail of the distribution

Daudin Etienne and Vallois have obtained another approximation for the tail of the distribution, i.e. if a is large enough:

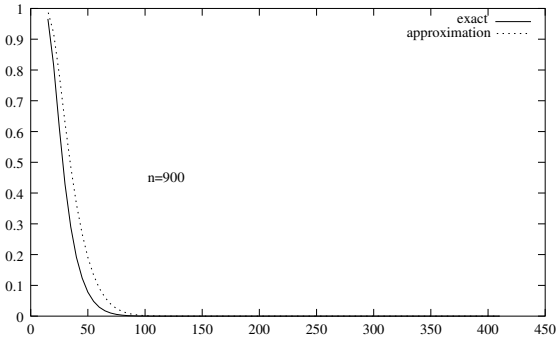

Figure 2.17: Centered Brownian approximation,$\mathbb{P}(X = -1) = 0.51$, $\mathbb{P}(X = 1) = 0.49$, $n = 900$, $\delta_{900} = -0.6$

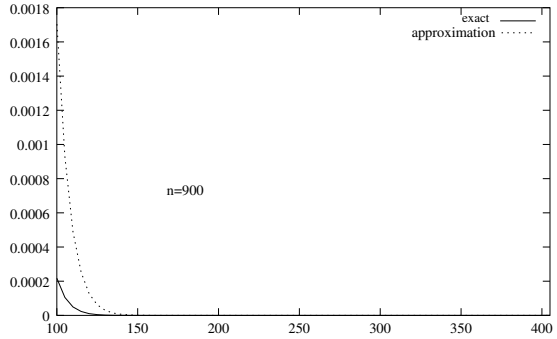

Figure 2.18: Centered Brownian approximation, Zoom on figure 2.17

$$\mathbb{P}\left(H_n \geq a\right) \underset{a\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}}\,\frac{\sigma\sqrt{n}}{a}\ \exp-\left(\delta_n^2/(2\sigma^2)\right)\ \exp\left(\frac{\delta_n a}{\sigma^2\sqrt{n}} - \frac{a^2}{2\sigma^2 n}\right). \tag{2.3.3}$$

This approximation is only available for large values of a. Some numerical tests seem to prove a good behaviour for a larger than $a_{inf} = \sqrt{n}(\mid \delta_n \mid + \sqrt{\delta_n^2 + 6\sigma^2})$. We investigate the properties of this approximation in two cases: $E(X) = 0.01$ and $E(X) = -0.02$ still using a Bernoulli p.d.f.

Bernoulli variables with $E(X) = 0.01$

The first example concerns the law of Bernoulli with following p.d.f. :

$$\mathbb{P}\left(X_i = 1\right) = 0.505, \quad \mathbb{P}\left(X_i = -1\right) = 0.495, \quad \mathbb{E}\left(X_i\right) = 0.01$$

$$\sigma = 9.99949e-01, \quad \delta_n = 5.00025e-01$$

Figure 2.19 shows the global behaviour of the exact distribution and the approximation (dotted line).

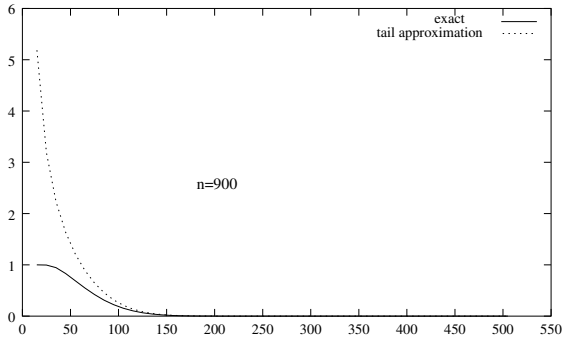

Figure 2.19: Non-centered Brownian approximation,$\mathbb{P}(X = 1) = 0.505$, $\mathbb{P}(X = -1) = 0.495$, $E(X) = 0.01$, $n = 900$, $\delta_{900} = 0.3$

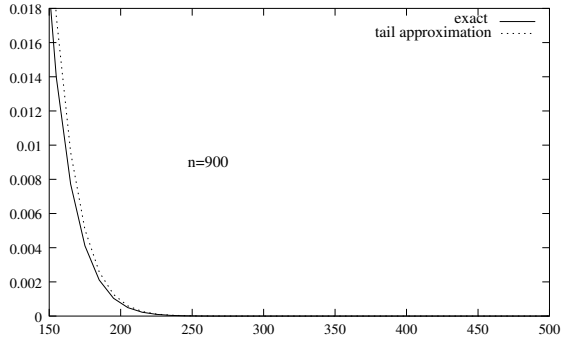

Figure 2.20: Non-centered Brownian approximation,Zoom on figure 2.19

A zoom on the preceding figure after $a_{inf} = 152$ gives the figure 2.20

The figures 2.19, 2.20 show a good behaviour of this approximation for $a > 150$.

Bernoulli variables with $E(X) = -0.02$

The p.d.f. of X is :

$$\mathbb{P}(X_i = 1) = 0.49, \quad \mathbb{P}(X_i = -1) = 0.51, \quad \mathbb{E}(X_i) = -0.02$$

$$\sigma = 9.99949e - 01, \quad \delta_n = -0.6$$

We don't examine the behaviour of this approximation for smallest value of $E(X)$ because we have noted that in this case Brownian approximation is valid and these two approximations are very close for small values of δ_n. The beginning of the domain of validity for this approximation is $a_{inf} = 30 * (0.6 + \sqrt{-0.6 + 6} = 87.7$.

This approximation is interesting because its scope of validity is large provided that a is sufficiently high. But when the parameter δ_n is too high,

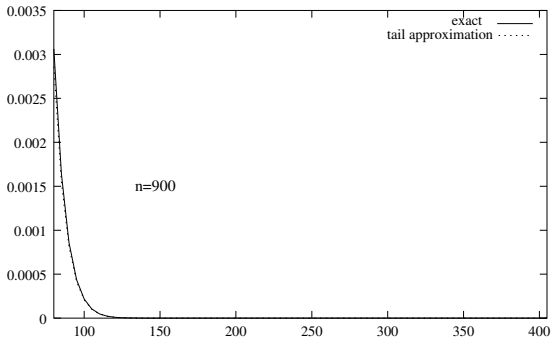

Figure 2.21: Non-centered Brownian approximation,$\mathbb{P}(X = 1) = 0.49$, $\mathbb{P}(X = -1) = 0.51$, $E(X) = -0.02$, $n = 900$, $\delta_{900} = -0.6$.

the values of a_{inf} are too high to be of practical interest. For instance for a Bernoulli distribution with $E(X) = -0.2$, $a_{inf} = 350$, but the probability for the local score with $n = 900$ variables to exceed 350 is equal to $2e - 70$ which is too small to be of practical interest.

2.4 Conclusion

We recall that using approximation (2.1.3) when $E(X)$ is near 0 may be quite misleading. In practice it is not recommended to use it when $E(X) > -0.2$. The centered Brownian approximation (2.1.5) should be used only if $|\delta_n| < 0.1$. The non centered Brownian approximation (2.1.6) may be used on a larger scope provided a is sufficiently large. We can summarize our conclusions by the Figure 2.22 which gives the scope of validity of each approximation.

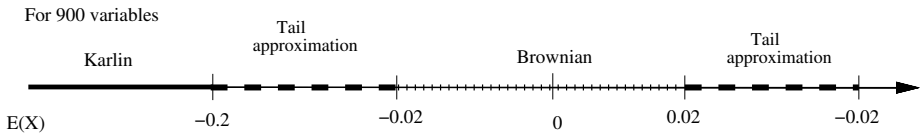

Figure 2.22: Scope of validity of each approximation in function of $E(X)$

Bibliography

[BDMR00] J.N. Bacro, J.J. Daudin, S. Mercier, and S. Robin. Back to the local score in the logarithmic case : a direct and simple proof. *to appear*, 2000.

[DEV00] J.J. Daudin, M.P. Etienne, and P. Vallois. Asymptotic behaviour of the local score of independant and identically distributed random sequence. *not published*, 2000.

[DK92] A. Dembo and S. Karlin. Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Prob*, 24:113–140, 1992. USA.

[DM99] J.J. Daudin and S. Mercier. Distribution exacte du score local d'une suite de variables indépendantes et identiquement distribuées. *C. R. Acad. Sciences*, 9:815–820, 1999. Série I, Math.

[KA90] S. Karlin and Altschul. Methods for assessing the statistical signifiance of molecular sequence features by using general scoring schemes. *Proc. Natl . Acad. Sci*, 87:2264–2268, 1990. USA.

Chapitre 3

Approximation of the distribution of the supremum of a centred random walk. Application to the local score

Approximation of the distribution of the supremum of a centred random walk. Application to the local score.

Marie Pierre ETIENNE[a], Pierre VALLOIS[a].

[a]Institut de Mathématiques Elie Cartan, Université Henri Poincaré.
BP. 239, 54506 Vandoeuvre Lès Nancy Cedex, France.
E-mail : Marie-Pierre.Etienne@iecn.u-nancy.fr
E-mail : Pierre.Vallois@iecn.u-nancy.fr

Abstract

Let $(X_n)_{n\geq 0}$ be a real random walk starting at 0, with increments bounded by a constant K. The main result of this study is : $|\mathbb{P}\left(\frac{S_n}{\sqrt{n}} \geq x\right) - \mathbb{P}\left(M_1 \geq x\right)| \leq C(n,K)\sqrt{\frac{\ln n}{n}}$, *where $x \geq 0$, S_n (respectively M_1) is the supremum at time n (resp. 1) of the random walk (resp. a linear Brownian motion) and $C(n,K)$ is an explicit constant. We also prove that in the previous inequality S_n can be replaced by the local score.*

Keywords : Skorokhod's embedding, random walk, local score, maximum.

AMS 2000 Subject classifications
60F05, 60F17, 60G17, 60G40, 60G50, 60J65.

3.1 Introduction

Let $(\xi_i)_{i\geq 1}$ be a sequence of i.i.d. random variables, with zero mean and variance σ^2. We denote by $(X_n)_{n\geq 0}$ the associated random walk :

$$X_0 = 0, \qquad X_n = \sum_{i=1}^{n} \xi_i, \qquad n \geq 1. \tag{3.1.1}$$

1) The well known central limit theorem (CLT) tells us that for every x in $\mathbb{R}$, $\lim_{n\to\infty} \mathbb{P}\left(\frac{X_n}{\sigma\sqrt{n}} \geq x\right) = \mathbb{P}\left(G \geq x\right)$ where G is a $\mathcal{N}(0,1)$-Gaussian

random variable. In practice it is often important to estimate the rate of convergence. Loëve ([Bil68] and [Loè79] p.288) has proved :

$$| \mathbb{P}\left(\frac{X_n}{\sigma\sqrt{n}} \geq x\right) - \mathbb{P}\left(G \geq x\right) | \leq \frac{C\mathbb{E}\left[|\xi_1|^3\right]}{\sqrt{n}}; \quad x \in \mathbb{R}, n \geq 1; \qquad (3.1.2)$$

where C is a constant.
2) Suppose now that we are interested in the asymptotic behaviour of S_n, as n goes to infinity, $S_n = \max_{0\leq i\leq n} X_i$. The CLT is not sufficient, we need a functional convergence result (Donsker's theorem [Bil68] p.68), which implies :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \geq x\right) = \mathbb{P}\left(M_1 \geq x\right); \quad x \geq 0, \qquad (3.1.3)$$

where $M_1 = \sup_{0\leq t\leq 1} B_t$, and $(B_t, t \geq 0)$ is a standard one dimensional Brownian motion started at 0.
Since M_1 and $|B_1|$ are identically distributed, the right hand-side of (3.1.3) can be easily computed.
A priori the rate of convergence of $\mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \geq x\right)$ to $\mathbb{P}\left(M_1 \geq x\right)$ is unknown.
3) In [DEV00], motivated by biological considerations, we established a similar result to (3.1.3) :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{H_n}{\sigma\sqrt{n}} \geq x\right) = \mathbb{P}\left(B_1^* \geq x\right); \quad x \geq 0, \qquad (3.1.4)$$

where $H_n = \max_{0\leq i\leq j} \left(X_j - X_i\right)$ and $B_1^* = \sup_{0\leq t\leq 1} |B_t|$. Recall that the density function of B_1^* can be expressed through series (cf [BS96],p.146 and annex A in [DEV00]).
The analysis of genetic sequences requires a precise estimate of $\mathbb{P}\left(\frac{H_n}{\sigma\sqrt{n}} \geq x\right)$. However the rate of decay of $n \to |\mathbb{P}\left(\frac{H_n}{\sigma\sqrt{n}} \geq x\right) - \mathbb{P}\left(B_1^* \geq x\right)|$ is unknown. Therefore its knowledge would be useful.

4) The aim of this work is to give effective bounds to

$$\delta_n(S) = | \mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \geq x\right) - \mathbb{P}\left(M_1 \geq x\right) |$$

and to

$$\delta_n(H) = | \mathbb{P}\left(\frac{H_n}{\sigma\sqrt{n}} \geq x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} |B_1| \geq x\right) |.$$

We prove (cf theorems 20 and 27) the following inequality :

$$\delta_n(Z) \leq C\sqrt{\frac{\ln n}{n}},$$

where $Z = S$ or H and C is a computable constant which only depends of the law of (ξ_i).

Let us detail the organization of the paper. In section 3.2 we deal with the supremum of a centred random walk. The previous analysis can be adapted (cf section 3.3) to handle the local score and is presented in section 3.3. In section 3.3.2, with the process S, we check the accuracy of previous bounds through numerical tests.

3.2 Approximation of the distribution of the supremum

1) Let $(\xi_i)_{i\geq 1}$ be a sequence of i.i.d. bounded random variables with 0 mean. We set

$$X_0 = 0, \quad X_n = \sum_{i=1}^{n} \xi_i, \quad n \geq 1. \tag{3.2.1}$$

We denote by σ^2 the variance of ξ_i and we assume :

$$|\xi_n| \leq K, \quad \forall n \geq 1. \tag{3.2.2}$$

The main idea of our approach is to embed the random walk $(X_n)_{n\geq 0}$ in a Brownian motion. The random walk $(X_n)_{n\geq 0}$ can be actually considered as a Brownian motion stopped at an increasing sequence of stopping times.
We recall below the scheme introduced by Skorokhod [Sko65] which allows to represent the random walk $(X_n)_{n\geq 0}$ as $(B_{T_n}, n \geq 0)$, where $(B_t, t \geq 0)$ is a standard one dimensional Brownian motion started at 0, and $(T_n)_{n\geq 0}$ is an increasing sequence of stopping times. This representation is the key of our approach.

2) If μ is a probability measure on $\mathbb{R}$ centred and having a finite first moment (i.e $\int_{\mathbb{R}} |x|\mu(\mathrm{d}x) < +\infty$ and $\int_{\mathbb{R}} x\mu(\mathrm{d}x) = 0$) we know ([AY79] and [Val83]) that there exists a stopping time T such that

$$\text{the law of } B_T \text{ is } \mu, \tag{3.2.3}$$

and

$$(B_{T\wedge t}, t \geq 0) \text{ is a uniformly integrable martingale.} \tag{3.2.4}$$

(3.2.4) tells us that T can be chosen not too large.
In fact if μ has a compact support included in $[-A, A]$, maximal inequality and (3.2.4) imply :

$$T \leq T^*(A), \tag{3.2.5}$$

where $T^*(A) = \inf\{t \geq 0, |B_t| \geq A\}$.
Conversely (3.2.5) implies (3.2.4).
In our approach we only deal with random walk having bounded increments. Then we restrict ourself to probability measures with compact support, or Brownian stopping time verifying (3.2.5).

Let $\mathcal{P}_c$ the set of probability measures on $\mathbb{R}$, with compact support and centred. We denote by $(U(\mu))_{\mu\in\mathcal{P}_c}$ a family of stopping times such that :

$$B_{U(\mu)} \sim \mu, \quad Supp(\mu) \subset [-K, K], \quad U(\mu) \leq T^*(K). \tag{3.2.6}$$

We need a little bit more than (3.2.6), we assume $\mathcal{P}_c$ has the following scaling property :

$$U(\mu_c) \overset{(d)}{=} c^2 U(\mu), \text{ for any } c > 0, \tag{3.2.7}$$

where μ_c is the image of μ by $x \mapsto c\,x$.
The two families of stopping times defined by [AY79] and [Val83] verify these properties.

We are now able to state the main result of this section, concerning the asymptotic behaviour of S_M, as M goes to infinity, where $S_k = \max_{0\leq i\leq k} X_i$.

Theorem 20 *For all $x \geq 0$ and $M \geq 10$:*

$$|\mathbb{P}\left(\frac{S_M}{\sqrt{M}} \geq x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right)| \leq C(M,\mu)\sqrt{\frac{\ln M}{M}}. \tag{3.2.8}$$

where

$$C(M,\mu) = \frac{2K}{\sigma\sqrt{2\pi}}\frac{1}{\sqrt{\ln M}} + \frac{1}{\sqrt{\ln M}} + \frac{2\,e^{-1/2}\sqrt{\mathbb{E}(U(\mu)^2) - \sigma^4}}{\sqrt{\pi}\sigma^2}. \tag{3.2.9}$$

Moreover $C(M,\mu) \leq \hat{C}(M, K/\sigma)$, where

$$\hat{C}(M,y) \leq \frac{y}{\sqrt{\pi}} + \frac{1}{\sqrt{2}} + 2\,e^{-1/2}\sqrt{\frac{\frac{5y^4}{3} - 1}{\pi}}. \tag{3.2.10}$$

3) Obviously (3.2.8) is meaningful if M is large. However as it is said in the introduction we want to have explicit bounds, M being fixed.
In the sequel M is a scale parameter, M being an integer larger than 1.
We presently give a representation of the random walk $(X_k)_{k\geq 0}$ in terms of Brownian motion path.

Proposition 21 *There exists a sequence of stopping times $(T_n)_{n\geq 0}$, such that :*

$$T_0 = 0, \quad T_k = \sum_{1\leq i\leq k} T_i', \tag{3.2.11}$$

and

$$(\sigma B_{T_k}, k \geq 0) \overset{(d)}{=} \left(\frac{X_k}{\sqrt{M}}, k \geq 0\right), \tag{3.2.12}$$

where $(T_i')_{i\geq 1}$ are independent random variables, each T_i' belonging to $U(\nu)$, ν being the common distribution of $\xi/\sigma\sqrt{M}$ In particular :

$$B_{T_i'} \stackrel{(d)}{=} \frac{\xi_i}{\sigma\sqrt{M}}.$$

Proof : We set $T_1 = U(\nu)$. Property (3.2.6) implies that $B_{T_1} \stackrel{(d)}{=} X_1/\sigma\sqrt{M} = \xi_1/\sigma\sqrt{M}$.
We know that $(B_t' = B_{t+T_1} - B_{T_1}, t \geq 0)$ is a one dimensional Brownian motion, independent of B_{T_1}. Let T_2' be a stopping time $U'(\nu)$ (associated with ν and $(B_t'; t \geq 0)$) such that $B'_{T_2'} \stackrel{(d)}{=} \xi_2/\sigma\sqrt{M}$, and

$$T_2' \leq \inf\left\{t \geq 0, |B_t'| \geq \frac{K}{\sigma\sqrt{M}}\right\}.$$

Iterating this procedure, we define by induction an increasing sequence of random times $(T_k, k \geq 0)$ such that :

$$T_1' = T_1, \tag{3.2.13}$$

$$B_{T_k+T'_{k+1}} - B_{T_k} = B_{T_{k+1}} - B_{T_k} \stackrel{(d)}{=} \frac{1}{\sigma\sqrt{M}}\xi_{k+1}; \quad \forall k \geq 0, \tag{3.2.14}$$

where

$$T_0 = 0, \quad T_k = T_1' + \ldots + T_k'; \quad k \geq 1. \tag{3.2.15}$$

T'_{k+1} is a stopping time with respect to the filtration generated by the Brownian motion $(B_{T_k+t} - B_{T_k}; t \geq 0)$. In particular

$$(B_{T_k} - B_{T_{k-1}}; k \geq 1) \stackrel{(d)}{=} \left(\frac{\xi_k}{\sigma\sqrt{M}}; k \geq 1\right). \tag{3.2.16}$$

□

In our study we are looking for properties of the law of $S_M = \max_{0\leq i\leq M} X_i$. Obviously it depends only on the law of the whole process $(X_k)_{k\geq 0}$. Therefore we can choose any realization of the random walk $(X_k)_{k\geq 0}$. In the sequel of the paper, according to proposition 21, we take :

$$X_k = \sigma\sqrt{M}B_{T_k}, \quad \forall\, k \geq 1. \tag{3.2.17}$$

We use the strength of (3.2.17) to obtain first bounds to $\mathbb{P}\left(S_M/\sqrt{M} \geq x\right)$. The key point of our method is the following lemma :

Lemma 22 *We have :*

$$\frac{1}{\sqrt{M}}S_k \leq \sigma \sup_{0\leq u\leq T_k} B_u \leq \frac{1}{\sqrt{M}}S_k + \frac{K}{\sqrt{M}}; \qquad \forall k \geq 1, \tag{3.2.18}$$

$$\mathbb{P}\left(\sup_{0\le u\le 1} B_u \ge \frac{1}{\sigma\sqrt{1-\varepsilon}}\left(x+\frac{K}{\sqrt{M}}\right)\right)-\mathbb{P}\left(|T_M-1|\ge\varepsilon\right)\le \mathbb{P}\left(\frac{S_M}{\sqrt{M}}\ge x\right), \tag{3.2.19}$$

$$\mathbb{P}\left(\frac{S_M}{\sqrt{M}}\ge x\right)\le \mathbb{P}\left(\sup_{0\le u\le 1} B_u \ge \frac{x}{\sigma\sqrt{1+\varepsilon}}\right)+\mathbb{P}\left(|T_M-1|\ge\varepsilon\right), \tag{3.2.20}$$

for any $x\ge 0$ *and* $\varepsilon>0$.

Proof : a) (3.2.17) implies (3.2.18).
b) Let $\varepsilon>0$ and $x\ge 0$. The first inequality in (3.2.18) implies :

$$\mathbb{P}\left(\frac{S_M}{\sqrt{M}}\ge x\right)\le \mathbb{P}\left(\sup_{0\le u\le T_M} B_u \ge \frac{x}{\sigma}\right).$$

We decompose the probability in the right hand-side as follow :

$$\begin{aligned}\mathbb{P}\left(\sup_{0\le u\le T_M} B_u \ge x/\sigma\right) &\le \mathbb{P}\left(|T_M-1|\ge\varepsilon\right)\\ &\quad+\mathbb{P}\left(T_M\le 1+\varepsilon,\ \sup_{0\le u\le T_M} B_u\ge x/\sigma\right),\\ &\le \mathbb{P}\left(|T_M-1|\ge\varepsilon\right)+\mathbb{P}\left(\sup_{0\le u\le 1+\varepsilon} B_u\ge x/\sigma\right).\end{aligned}$$

Since the Brownian motion $(B_t, t\ge 0)$ has the scaling property :

$$(B_{tc}, t\ge 0)\overset{(d)}{=}(\sqrt{c}B_t, t\ge 0)$$

for any $c>0$,

$$\sup_{0\le u\le c} B_u \overset{(d)}{=} \sqrt{c}\sup_{0\le u\le 1} B_u.$$

This achieves the proof of (3.2.20).
c) (3.2.19) is a direct consequence of the following inclusions :

$$\begin{aligned}&\left\{\sup_{0\le u\le 1-\varepsilon} B_u\ge\frac{x}{\sigma}+\frac{K}{\sigma\sqrt{M}}, |T_M-1|\le\varepsilon\right\}\\ &\quad\subset\left\{\sup_{0\le u\le T_M} B_u\ge\frac{x}{\sigma}+\frac{K}{\sigma\sqrt{M}}, |T_M-1|\le\varepsilon\right\}\\ &\quad\subset\left\{\frac{S_M}{\sqrt{M}}\ge x, |T_M-1|\le\varepsilon\right\}\subset\left\{\frac{S_M}{\sqrt{M}}\ge x\right\}.\end{aligned}$$

□

We note that (3.2.15) implies that

$$\mathbb{E}(T_M) = M\mathbb{E}(T_1) = M\mathbb{E}\left(B^2_{T_1}\right) = \frac{M}{\sigma^2 M}\mathbb{E}\left(\xi_1^2\right) = 1.$$

Moreover $T_M = T'_1 + \ldots + T'_M$, and $(T'_i)_{1\leq i\leq M}$ are i.i.d., then the weak law of large numbers implies that T_M converges to 1, in probability, as M goes to infinity. Consequently $\lim_{M\to\infty}\mathbb{P}(|T_M - 1| \geq \varepsilon) = 0$.

Recall that our goal is to look for effective bounds for $\mathbb{P}\left(S_M/\sqrt{M} \geq x\right)$, x and M being given.
This leads us to take ε as a function of M in order to minimize $\mathbb{P}(|T_M - 1| \geq \varepsilon)$. This can be done through a large deviation technique, because the stopping time $T^*(A)$ admits some small exponential moments. Since for every probability measure μ with compact support in $[-K, K]$ we have $U(\mu) \leq T^*(K)$, there exists $A(\mu) > 0$ such that :

$$\mathbb{E}[\exp\{\lambda U(\mu)\}] < +\infty \Leftrightarrow \lambda < A(\mu). \tag{3.2.21}$$

Lemma 23 *Let $M \geq 1$ and $\varepsilon \geq 1$. We assume that μ is centred and has a compact support, recall that μ is the common law of (ξ_i). Then for any $\lambda_1 \in [0, A(\mu)[$, $\lambda_2 > 0$, we have :*

$$\mathbb{P}(T_M - 1 \geq \varepsilon) \leq \exp\{-Mf_\varepsilon(\lambda_1)\}, \tag{3.2.22}$$

$$\mathbb{P}(T_M - 1 \leq -\varepsilon) \leq \exp\{-Mg_\varepsilon(\lambda_2)\}, \tag{3.2.23}$$

where

$$f_\varepsilon(x) = \sigma^2(1+\varepsilon)x - \ln(\mathbb{E}[\exp(xU(\mu))]), \quad x < A(\mu), \tag{3.2.24}$$

and

$$g_\varepsilon(x) = -\sigma^2(1-\varepsilon)x - \ln(\mathbb{E}[\exp(-xU(\mu))]), \quad x \geq 0. \tag{3.2.25}$$

Proof : The crucial identity is :

$$T_M = T'_1 + \ldots + T'_M.$$

Recall that $(T'_i)_{1\leq i\leq M}$ are independent and distributed as $T'_1 = T_1$.
1) Let $\lambda > 0$. Then, using Markov's inequality

$$\begin{aligned}\mathbb{P}(T_M \geq 1+\varepsilon) &= \mathbb{P}\left(\exp\left\{\lambda\left(T'_1 + \ldots + T'_M\right)\right\} \geq \exp\{\lambda(1+\varepsilon)\}\right)\\ &\leq \exp -\lambda(1+\varepsilon)\left(\mathbb{E}\left[e^{\lambda T_1}\right]\right)^M. \end{aligned}\tag{3.2.26}$$

T_1 is a stopping time associated with the distribution of $\xi_1/\sigma\sqrt{M}$, so

$$T_1 = U(\mu_c) \quad \text{where } c = \frac{1}{\sigma\sqrt{M}}.$$

Using the scaling property (3.2.7) :

$$\mathbb{E}\left[e^{\lambda T_1}\right] = \mathbb{E}\left[\exp\left\{\frac{\lambda}{\sigma^2 M}U(\mu)\right\}\right].$$

Then

$$\mathbb{P}\left(T_M \geq 1+\varepsilon\right) \leq \exp\left\{-M\left(\frac{\lambda}{M}(1+\varepsilon) - \ln\left(\mathbb{E}\left[\exp\left\{\frac{\lambda}{\sigma^2 M}U(\mu)\right\}\right]\right)\right)\right\}.$$

(3.2.22) follows immediately.

2) As for (3.2.23) it is sufficient to replace (3.2.26) by :

$$\mathbb{P}\left(T_M \leq 1-\varepsilon\right) = \mathbb{P}\left(\exp\left\{-\lambda\left(T_1' + \ldots + T_M'\right)\right\} \geq \exp\left\{-\lambda(1-\varepsilon)\right\}\right).$$

□

Lemma 24 *There exists* $0 < A' \leq A(\mu)$ *such that for any* $\varepsilon \leq \frac{\mathbb{E}[U(\mu)^2]-\sigma^4}{\sigma^2}A'$,

$$\mathbb{P}\left(|T_M - 1| \geq \varepsilon\right) \leq 2\exp\left(-c_1(\mu)M\varepsilon^2\right), \tag{3.2.27}$$

where

$$c_1(\mu) = \frac{\sigma^4}{4\left(\mathbb{E}(U(\mu)^2) - \sigma^4\right)} > 0. \tag{3.2.28}$$

Proof : 1) According to lemma 23, the search of an upper bound for $\mathbb{P}\left(T_M - 1 \geq \varepsilon\right)$ leads us to study f_ε. At this step, ε and μ are fixed, f stands for f_ε and $U(\mu)$ will be designed by U.
2) We have

$$\begin{aligned} L_\mu(x) &= \mathbb{E}\left[\exp xU\right] = 1 + x\mathbb{E}\left[U\right] + \frac{x^2}{2}\mathbb{E}\left[U^2\right] + o(x^2), \\ &= 1 + x\sigma^2 + \frac{x^2}{2}\mathbb{E}\left[U^2\right] + o(x^2). \end{aligned}$$

Then

$$\ln\left(L_\mu(x)\right) = x\sigma^2 + \frac{x^2}{2}\mathbb{E}\left[U^2\right] - \frac{x^2\sigma^4}{2} + o(x^2),$$

and

$$\begin{aligned} f(x) &= \sigma^2(1+\varepsilon)x - x\sigma^2 - \frac{x^2}{2}\left(\mathbb{E}\left[U^2\right] - \sigma^4\right) + o(x^2), \\ &= h(x) + o(x^2). \end{aligned}$$

where

$$h(x) = \sigma^2\epsilon x - \frac{x^2}{2}\left(\mathbb{E}\left[U^2\right] - \sigma^4\right).$$

Consequently there exists $0 < A' < A(\mu)$ such that

$$f(x) \geq h(x) - \frac{\mathbb{E}\left[U^2\right] - \sigma^4}{4} x^2, \qquad \forall x \in [0, A']. \tag{3.2.29}$$

Let us remark that $\mathbb{E}\left[U\right]^2 = \sigma^4 \leq \mathbb{E}\left[U^2\right]$. Then h admits a maximum at point x_* :

$$x_* = \frac{\sigma^2 \varepsilon}{\mathbb{E}\left[U^2\right] - \sigma^4}.$$

Using (3.2.29), we obtain :

$$f(x_*) \geq \frac{\sigma^4}{4\left(\mathbb{E}\left[U^2\right] - \sigma^4\right)} \varepsilon^2, \text{ as soon as } x_* \leq A'.$$

Since $x_* < A' \Leftrightarrow \varepsilon \leq \frac{\mathbb{E}\left[U^2\right] - \sigma^4}{\sigma^2} A'$, thus

$$\mathbb{P}\left(T_M - 1 \geq \varepsilon\right) \leq \exp\left\{-c_1(\mu) M \varepsilon\right\}, \qquad \forall \varepsilon \leq \frac{\mathbb{E}\left[U^2\right] - \sigma^4}{\sigma^2} A', \tag{3.2.30}$$

where

$$c_1(\mu) = \frac{\sigma^4}{4\left(\mathbb{E}\left[U^2\right] - \sigma^4\right)}.$$

3) We will now study g_ε. Since $x \mapsto \ln \mathbb{E}\left[-xU\right]$ is a convex function, with similar arguments as for f, replacing x by $-x$,

$$\ln \mathbb{E}\left[\exp\left\{-xU\right\}\right] = -x\sigma^2 + \frac{x^2}{2}\mathbb{E}\left[U^2\right] - \frac{x^2\sigma^4}{2} + o(x^2),$$

$$g_\varepsilon(x) = h(x) + o(x^2).$$

By the same way as previously, we have

$$\mathbb{P}\left(T_M - 1 \leq -\varepsilon\right) \leq \exp\left\{-c_2(\mu) M \varepsilon\right\}, \qquad \forall \varepsilon \leq \frac{\mathbb{E}\left[U^2\right] - \sigma^4}{\sigma^2} A'. \tag{3.2.31}$$

□

Lemma 25 *For any* $0 < \varepsilon < 1/2$,

$$\mathbb{P}\left(\sup_{0 \leq u \leq 1} B_u \geq \frac{x}{\sigma\sqrt{1+\varepsilon}}\right) \leq c\,\varepsilon + \mathbb{P}\left(\sup_{0 \leq u \leq 1} B_u \geq \frac{x}{\sigma}\right), \tag{3.2.32}$$

where

$$c = \frac{1}{2}\sqrt{\frac{3}{\pi}}\, e^{-1/2}.$$

Proof : As $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{0\le u\le 1} B_u \ge \frac{x}{\sigma\sqrt{1+\varepsilon}}\right) = \mathbb{P}\left(\sup_{0\le u\le 1} B_u \ge \frac{x}{\sigma}\right) + \delta,$$

where

$$\delta = \mathbb{P}\left(\frac{x}{\sigma\sqrt{1+\varepsilon}} \le \sup_{0\le u\le 1} B_u \le \frac{x}{\sigma}\right).$$

But it is well known that $\sup_{0\le u\le 1} B_u \stackrel{(d)}{=} |B_1|$, so that :

$$\begin{aligned}\delta &= \mathbb{P}\left(\frac{x}{\sigma\sqrt{1+\varepsilon}} \le |B_1| \le \frac{x}{\sigma}\right) = 2\,\mathbb{P}\left(\frac{x}{\sigma\sqrt{1+\varepsilon}} \le B_1 \le \frac{x}{\sigma}\right)\\ &= 2\left(\Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma\sqrt{1+\varepsilon}}\right)\right),\end{aligned}$$

with

$$\Phi(z) = \mathbb{P}\left(B_1 \le z\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-u^2/2}\mathrm{d}u.$$

Using formula of finite increments, we obtain :

$$\delta = 2\left(\frac{x}{\sigma} - \frac{x}{\sigma\sqrt{1+\varepsilon}}\right)\Phi'(y), \qquad \text{for some } y \in \left[\frac{x}{\sigma\sqrt{1+\varepsilon}}; \frac{x}{\sigma}\right].$$

However

$$0 < \frac{x}{\sigma} - \frac{x}{\sigma\sqrt{1+\varepsilon}} = \frac{x\varepsilon}{\sigma\sqrt{1+\varepsilon}\left(\sqrt{1+\varepsilon}+1\right)} \le \frac{x\varepsilon}{2\sigma}.$$

Suppose that $\varepsilon < 1/2$ and $y \in \left[\frac{x}{\sigma\sqrt{1+\varepsilon}}, \frac{x}{\sigma}\right]$, then

$$\Phi'(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2} \le \frac{1}{\sqrt{2\pi}}e^{-x^2/(3\sigma^2)}.$$

So that

$$\delta \le \varepsilon h_0\left(\frac{x}{\sigma}\right),$$

where

$$h_0(z) = \frac{z}{\sqrt{2\pi}}e^{-z^2/3}.$$

But $h_0(z) \le h_0(\sqrt{3/2}) = c$, this shows (3.2.32). □

At this stage we have to give a lower bound to $\mathbb{P}\left(\sup_{0\le u\le 1} B_u \ge \frac{1}{\sigma\sqrt{1-\varepsilon}}\left(x + \frac{K}{\sigma\sqrt{M}}\right)\right)$. Using same tools as for lemma 25, we will prove :

Lemma 26 *For any* $0 < \varepsilon < 1/2$,

$$\mathbb{P}\left(\sup_{0\le u\le 1} B_u \ge \frac{x}{\sigma}\right) \quad - \quad \frac{2K}{\sigma\sqrt{2\pi M}} - c_2\,\varepsilon$$

$$\leq \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{1}{\sigma\sqrt{1-\varepsilon}}\left(x+\frac{K}{\sqrt{M}}\right)\right) \quad (3.2.33)$$

where

$$c_2 = \frac{2e^{-1/2}}{\sqrt{2\pi}}.$$

Proof : 1)We set $y = x + K/\sqrt{M}$. Using the same arguments as for lemma 25, we obtain :

$$\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{y}{\sigma\sqrt{1-\varepsilon}}\right) = \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{y}{\sigma}\right) - \delta,$$

where

$$\delta = \mathbb{P}\left(\frac{y}{\sigma} \leq |B_1| \leq \frac{y}{\sigma\sqrt{1-\varepsilon}}\right) = 2\left(\Phi\left(\frac{y}{\sigma\sqrt{1-\varepsilon}}\right) - \Phi\left(\frac{y}{\sigma}\right)\right).$$

We have successively :

$$\delta = 2\left(\frac{y}{\sigma\sqrt{1-\varepsilon}} - \frac{y}{\sigma}\right)\Phi'(z), \quad \text{for some } z \in \left[\frac{y}{\sigma}; \frac{y}{\sigma\sqrt{1-\varepsilon}}\right].$$

Since $z \geq y/\sigma$,

$$\Phi'(z) \leq \frac{1}{\sqrt{2\pi}} \exp -\frac{y^2}{2\sigma^2},$$

and

$$\frac{y}{\sigma\sqrt{1-\varepsilon}} - \frac{y}{\sigma} = \frac{y}{\sigma}\left(\frac{1-\sqrt{1-\varepsilon}}{\sqrt{1-\varepsilon}}\right) = \frac{y}{\sigma}\left(\frac{\varepsilon}{(1+\sqrt{1-\varepsilon})(\sqrt{1-\varepsilon})}\right).$$

$$\delta \leq \varepsilon \frac{1}{(1+\sqrt{1-\varepsilon})(\sqrt{1-\varepsilon})} h_1\left(\frac{y}{\sigma}\right), \quad \text{with } h_1(z) = \frac{2z}{\sqrt{2\pi}} e^{-z^2/2}.$$

but $\varepsilon \leq 1/2$, so that $\sqrt{1-\varepsilon} \geq 1/\sqrt{2}$, then

$$(1+\sqrt{1-\varepsilon})(\sqrt{1-\varepsilon}) \geq \frac{1}{\sqrt{2}}\left(1+\frac{1}{\sqrt{2}}\right) = \frac{\sqrt{2}+1}{2} \geq 1.$$

We get

$$\delta \leq \varepsilon h_1\left(\frac{y}{\sigma}\right) \leq \varepsilon h_1(1) = \varepsilon c_2.$$

2) We have to express $\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq y/\sigma\right)$ through $\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq x/\sigma\right)$.

$$\begin{aligned} \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq x/\sigma\right) &- \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq y/\sigma\right) \\ &= \mathbb{P}\left(x/\sigma \leq \sup_{0\leq u\leq 1} B_u \leq x/\sigma + K/(\sigma\sqrt{M})\right), \end{aligned}$$

$$\begin{aligned} &= 2\left(\Phi\left(\frac{x}{\sigma}+\frac{K}{\sigma\sqrt{M}}\right)-\Phi\left(\frac{x}{\sigma}\right)\right), \\ &\leq \frac{2K}{\sigma\sqrt{2\pi\,M}}\, e^{-x^2/(2\sigma^2)}, \\ &\leq \frac{2K}{\sigma\sqrt{2\pi\,M}}. \end{aligned}$$

This ends the proof. □

We are now able to prove theorem 20. We can control the rate of convergence of the two probability distributions functions.

Proof of theorem 20 : Using lemma 22, (3.2.27), (3.2.32) and (3.2.33), we obtain :

$$\begin{aligned} &|\mathbb{P}\left(\frac{S_M}{\sqrt{M}}\geq x\right)-\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u\geq\frac{x}{\sigma}\right)| \\ &\qquad\leq \max\left\{\frac{2K}{\sigma\sqrt{2\pi M}}+\frac{2e^{-1/2}}{\sqrt{2\pi}}\varepsilon+2\exp\left\{-c_1(\mu)M\varepsilon^2\right\},\right. \\ &\qquad\qquad\left.\frac{1}{2}\sqrt{\frac{3}{\pi}}e^{-1/2}\varepsilon+2\exp\left\{-c_1(\mu)M\varepsilon^2\right\}\right\}. \end{aligned}$$

We are lead to choose the best ε under the following assumption :

$$\varepsilon\leq\frac{\mathbb{E}(U(\mu)^2)-\sigma^4}{\sigma^2}A',\ \text{and}\ \ \varepsilon\leq 1/2.$$

Choosing $\varepsilon=\sqrt{\frac{\ln M}{2\,M\,c_1(\mu)}}$, we obtain :

$$\begin{aligned} &|\mathbb{P}\left(\frac{S_M}{\sqrt{M}}\geq x\right)-\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u\geq\frac{x}{\sigma}\right)| \\ &\qquad\leq \frac{2K}{\sigma\sqrt{2\pi M}}+\frac{2e^{-1/2}}{\sqrt{2\pi}}\sqrt{\frac{\ln M}{2\,M\,c_1(\mu)}}+\frac{1}{\sqrt{M}}, \\ &\qquad\leq \sqrt{\frac{\ln M}{M}}\left(\frac{2K}{\sqrt{\ln M}\sigma\sqrt{2\pi}}+\frac{1}{\sqrt{\ln M}}+\frac{e^{-1/2}}{\sqrt{\pi\,c_1(\mu)}}\right). \end{aligned}$$

As soon as M is more than 10, $\ln M$ is more than 2 and

$$\frac{2K}{\sqrt{\ln M}\sigma\sqrt{2\pi}}+\frac{1}{\sqrt{\ln M}}+\frac{e^{-1/2}}{\sqrt{\pi\,c_1(\mu)}}\leq\frac{K}{\sigma\sqrt{\pi}}+\frac{1}{\sqrt{2}}+\frac{e^{-1/2}}{\sqrt{\pi\,c_1(\mu)}}.$$

Since $c_1(\mu)$ is given by (3.2.28), (3.2.8) follows immediately.

The value of $\mathbb{E}\left[U(\mu)^2\right]$ depends on the choice of $U(\mu)$, but $U(\mu) \leq T_K^*$, so that

$$\mathbb{E}\left[U(\mu)^2\right] \leq \mathbb{E}\left[(T_K^*)^2\right] = K^4 \mathbb{E}\left[(T_1^*)^2\right] = \frac{5}{3} K^4.$$

□

3.3 Applications to the local score. Numerical tests.

If we replace $(X_n)_{n\geq 0}$ by $(-X_n)_{n\geq 0}$ in theorem 20 and we use the symmetry of Brownian motion (namely $(-B_t)_{t\geq 0} \overset{(d)}{=} (B_t)_{t\geq 0}$), we obtain without calculation :

$$|\mathbb{P}\left(\frac{\min_{0\leq i\leq M} X_i}{\sqrt{M}} \leq -x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right)| \leq \hat{C}(M, K/\sigma)\sqrt{\frac{\ln M}{M}},$$

$$\hat{C}(M, y) = \frac{y}{\sqrt{\pi}} + \frac{1}{\sqrt{2}} + 2\, e^{-1/2}\sqrt{\frac{\frac{5y^4}{3} - 1}{\pi}}.$$

Our scheme developed previously in section 3.2 is rich enough to be applied to the local score $(H_n)_{n\geq 0}$:

$$H_n = \max_{0\leq i\leq j\leq n} (X_j - X_i).$$

In the sequel we prove the analog of theorem 20 for the local score (theorem 27).
We also end up this paper with numerical computations.

3.3.1 The local score

As we did in section 3.2, we suppose that the random variables (ξ_i) are centred and bounded. Recall that $(X_n)_{n\geq 0}$ denotes the random walk associated to (ξ_i) (cf (3.1.1)). The local score H_M of $(X_n)_{n\geq 0}$ is

$$H_M = \max_{0\leq i\leq j\leq M} (X_j - X_i) = \max_{0\leq j\leq M} \left(X_j - \min_{0\leq i\leq j} X_i\right). \tag{3.3.1}$$

Let us state our main result involving the local score.

Theorem 27 *For all $x \geq 0$, $M \geq 10$,*

$$|\mathbb{P}\left(\frac{H_M}{\sqrt{M}} \geq x\right) - \mathbb{P}\left(\sigma \sup_{0\leq u\leq 1} |B_u| \geq x\right)| \leq C'(M, k/\sigma)\sqrt{\frac{\ln M}{M}}, \tag{3.3.2}$$

where

$$C'(M,y) = \frac{4y}{\sqrt{\pi}} + \frac{1}{\sqrt{2}} + 4e^{-1/2}\sqrt{\frac{\frac{5y^4}{3}-1}{\pi}}.$$

We are now able to prove the main result about the behaviour of the local score.

Proof of theorem 27 : The method is the same as the one developed for the maximum. However there are two changes.

a) (3.2.18) has to be replaced by :

$$\frac{1}{\sqrt{M}}H_k \leq \sigma \max_{0\leq u\leq T_k}\left(B_u - \min_{0\leq v\leq u} B_v\right) \leq \frac{1}{\sqrt{M}}H_k + \frac{2K}{\sqrt{M}}.$$

b) We need an upper-bound for $\mathbb{P}(a<\zeta<b)$, where $0<a<b$ and $\zeta = \max_{0\leq u\leq 1}(B_u - \min_{0\leq v\leq u} B_v)$. Recall that Lévy's theorem implies that $\zeta \overset{(d)}{=} B_1^*$, $B_1^* = \sup_{0\leq u\leq 1}|B_u|$.

If we set $S_1 = \sup_{0\leq u\leq 1} B_u$ and $I_1 = \min_{0\leq u\leq 1} B_u$, then

$$S_1 \overset{(d)}{=} -I_1 \overset{(d)}{=} |B_1|$$

and

$$\{a<B_1^*<b\} \subset \{a<S_1<b\} \cup \{a<-I_1<b\},$$
$$\mathbb{P}(a<B_1^*<b) \leq 2\mathbb{P}(a<|B_1|<b).$$

This allows us to reduce to the previous study dealing with the maximum. □

3.3.2 Numerical tests

We would like to verify the quality of our upper bound $C(M,\mu)$ in (3.2.9). Let μ be the probability measure, $\mu = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$ where δ is the Dirac measure.

Let us start with M fixed. We generate k times the sequence $(X_i)_{0\leq i\leq M}$ and then obtain a k-sample of $S_M/\sqrt{M}$ whose empirical distribution function is denoted $F_{k,M}$. Kolmogorov's theorem tells us that $\mathbb{P}\left(S_M/\sqrt{M}\geq x\right)$ can be approximated by $F_{k,M}(x)$, uniformly with respect to x, with heuristic rate $1/\sqrt{k}$. We choose $k = 100000$.

Recall that $\sup_{0\leq u\leq 1} B_u$ is distributed as $|B_1|$. We set

$$\delta_{k,M} = \sqrt{\frac{M}{\ln M}}\left(\sup_{x\in\mathbb{R}}|F_{k,M}(x) - \mathbb{P}(|B_1|\geq x/\sigma)|\right),$$

$$C(M,\mu) = \frac{2K}{\sigma\sqrt{\pi}}\frac{1}{\sqrt{\ln M}} + \frac{1}{\sqrt{\ln M}} + \frac{2e^{-1/2}\sqrt{\frac{5K^4}{3\sigma^4}-1}}{\sqrt{\pi}},$$

with $K = 1$ and $\sigma = 1$.
On the figure below we draw $M \to \delta_{k,M}$ and $M \to C(M, \mu)$, M varying from 10 to 1000. We test that theorem 20 holds : $\delta_{k,M} \leq C(M, \mu)$ and we observe that the ratio $C(M, \mu)/\delta_{k,M}$ is of order 8.
This means that the universal constant $C(M, \mu)$ is convenient and not too large. For different μ, we observe the same phenomenon : $\delta_{k,M}/C(M, \mu)$ is less than 10.

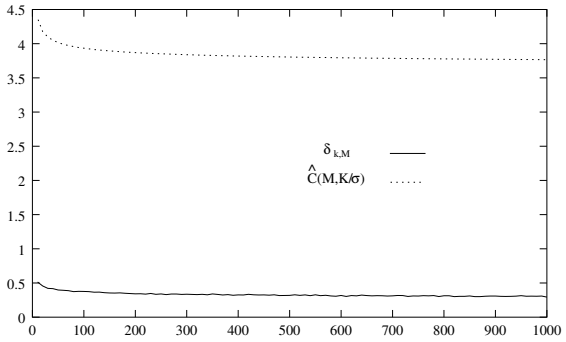

Figure 3.1: Symmetric Bernoulli law for ξ_i.

Bibliography

[AY79] Jacques Azema and Marc Yor. Une solution simple au probleme de Skorokhod. *Séminaire de probabilites XIII, Univ. Strasbourg 1977/78, Lect. Notes Math.*, 721:90–115, 1979. Seminaire de probabilites XIII, Univ. Strasbourg 1977/78.

[Bil68] P. Billingsley. *Convergence of probability measures.* John Wiley and Sons, 1968. New York.

[BS96] A. N. Borodin and P. Salminen. *Handbook of Brownian motion – Facts and formulae.* Birkhauser Verlag, 1996. Basel.

[DEV00] J.J. Daudin, M.P. Etienne, and P. Vallois. Asymptotic behaviour of the local score of independant and identically distributed random sequence. *not published*, 2000.

[Loè79] Michel Loève. *Probability theory.* Springer-Verlag, New York, fourth edition, 1979. Graduate Texts in Mathematics, Vol. 46.

[Sko65] Skorokhod, A.V. *Studies in the theory of random processes.* Reading, Mass.: Addison-Wesley Publish. Comp. Inc. VIII, 1965.

[Val83] Vallois, Pierre. Le probleme de Skorokhod sur R: Une approche avec le temps local. *Lect. Notes Math.*, 986:227–239, 1983. Sémin. de probabilites XVII, Proc. 1981/82,.

Troisième partie

Motifs communs à plus que deux séquences

Chapitre 1

L'alignement multiple

1.1 La recherche de motifs

On compare deux séquences entre elles lorsqu'on suppose, a priori, qu'elles ont des fonctions proches. Il est parfois utile de comparer tout un groupe de séquences entre elles pour en extraire une portion commune. En effet si parmi différentes espèces, on observe des protéines impliquées dans le même mécanisme, il est naturel de penser qu'elles ont un motif commun qui permet à chacune d'assurer sa fonction. Déterminer ce motif peut aider à la compréhension du mécanisme biologique mis en oeuvre.

alignement multiple par programmation dynamique des methodes d'alignement progressif

Des algos d'alignement multilpes.

rien n'empeche de definir un score d'alignement pour un groupement de sequences mais cet objet serait inutilisable en pratique.

des algos d'alignement multiple

1.2 Les difficultés liées à l'alignement multiple

Chapitre 2

Un algorithme de recherche de motifs

Convergence of a probabilistic algorithm to detect common patterns.

Madalina DEACONU[a], Marie-Pierre ETIENNE [a].

[a]Institut de Mathématiques Elie Cartan, Université Henri Poincaré.
BP. 239, 54506 Vandoeuvre Lès Nancy Cedex, France.
E-mail : Madalina.Deaconu@loria.fr
E-mail : Marie-Pierre.Etienne@iecn.u-nancy.fr

Abstract

Keywords :
AMS 1991 Subject classifications

2.1 Introduction

The problem of extracting common patterns from a set of sequences is very important in the study of DNA. Same patterns have mostly the same function or very closed function. Furthermore DNA segments which are conserved in different species could be indicated an important specific biological function.

The purpose of this paper is to give a probabilistic interpretation to a largely used biological algorithm proposed in [LSB$^+$93] which deals with the following problem : N sequences of length L composed of letters from a finite alphabet $\mathcal{A}$ are given. Its aim is to detect a common pattern of a given length W. The model assumes that a such pattern exists, we will prove the convergence of this algorithm.

The paper is divided in three parts. First we present and explain the algorithm. In a second part we construct the associated probabilistic model. finally we prove the convergence to a stationary measure.

2.2 Presentation of the algorithm

In order to detect common patterns in a set of sequences, traditionally we use a multiple alignment of sequences. The idea of the algorithm here is different. We consider one sequence at once, and we approximate the pattern by successive iterations of the algorithm.
We have first to precise our definition of a pattern of length W. In this work it designs a set of W independent probability distributions $(q_1, \ldots, q_W)$ on $\mathcal{A}$, where q_i is the probability distribution which governs the appearance of letters from $\mathcal{A}$ on the i-th position of the pattern. Obviously if the W distributions probability are all Dirac measures, the pattern we detect is an exact word.

Let us briefly explain the progress of the algorithm, section 2.2.3 will detail it, we just present the main steps here. It is initialized by choosing, uniformly in each sequence, the starting point of the pattern and it goes up in two steps.
1) First a sequence is chosen, uniformly, lets say i_0 and a model of the pattern is describe using (2.2.2).
2) the second steps chooses the starting point of the pattern using standard technic of maximum likelihood.

In order to detail the algorithm, we have to define the data and the variables of the problem.

2.2.1 The given information of the problem

- Sequences are stocked in a matrix A. $(A(i,j))_{\substack{i=1,\ldots,N \\ j=1,\ldots,L}}$, $A(i,j)$ is the value in the sequence i on the position j, $A(i,j) \in \{1, 2, 3, 4\}$.

- p is a vector of size 4, $p(i)$ notes the empirical frequency for the appearance of the letter i in the set of sequences. p is a vector of probabilities and its components are given by:

$$p(i) = \frac{\text{Number of appearance of } i}{NL}. \tag{2.2.1}$$

- The length W of the pattern is a given nonnegative integer number.

2.2.2 The variables

- x is a vector of size N. $x(i)$ notes the position of the beginning of the pattern in the sequence i.

- Q is a matrix of size $4 \times W$, $Q(l,k)$ notes the frequency for the coming up of the letter k on position l for the pattern, modified by a

perturbation. This value is $Q(l,k)$:

$$Q(l,k) = \frac{C_{lk} + p(k)\sqrt{N}}{N - 1 + \sqrt{N}}$$

where C_{lk} represents the number of appearances of the letter k on the position l of the pattern.

- Pos is a vector of size $L - W + 1$,

$$Pos(i) = \frac{\mathbb{P}(\text{pattern starts in } i)}{\mathbb{P}(\text{we are not in the pattern})}.$$

- *Choice* is an uniform random variable on $\{1, \ldots, N\}$ and represents the choice of the particular sequence on which we will shift the pattern.

2.2.3 The algorithm

The initialization step chooses, uniformly, in each sequence, the starting point of the pattern and this is done in an independent manner on the set of sequences. We stock these datum in a vector x. We evaluate also the probabilities $p(k)$ by applying formula (2.2.1).

The algorithm goes up in two steps.

1. We start by choosing uniformly a sequence, lets say i_0. We search now a pattern model by calculating $Q^{(i_0)}(l,k)$ for $k \in \{1, 2, 3, 4\}$ and $l \in \{1, \ldots, W\}$ with the formula:

$$Q^{(i_0)}(l,k) = \frac{C_{lk}^{(i_0)} + p(k)\sqrt{N}}{N - 1 + \sqrt{N}}, \tag{2.2.2}$$

   where $C_{lk}^{(i_0)}$ notes the number of coming up of letter k on position l of the pattern, in all sequences, excepting i_0.

2. The second step localizes the beginning of the pattern in the sequence i_0 l. We use for this the standard technic of maximum likelihood. In fact the pattern can start in the sequence i_0 on $L - W + 1$ possible places. We shall affect a weight $Pos(r)$ to each possible position r :

$$Pos(r) = \frac{Q(1, A(i_0, r))\ Q(2, A(i_0, r+1)) \ldots Q(W, A(i_0, r+W-1))}{p(A(i_0, r))p(A(i_0, r+1)) \ldots p(A(i_0, r+W-1))}. \tag{2.2.3}$$

   After, we choose the beginning of the pattern in the sequence i_0 according to the probability law :

$$\mathbb{P}\left(x(i_0) = r\right) = \frac{Pos(r)}{\sum_{s=1}^{L-W+1} Pos(s)}.$$

We iterate the steps 1 and 2 of this algorithm until ... (see the annex 2.4.1).

2.3 Markov chain associated with this algorithm

By looking to the steps of the algorithm it is clear that, from one step to another, the position of the beginning of the pattern depends only on the position of the beginning of the pattern at the previous step. A natural idea is thus to construct a Markov chain which describes the evolution of the position of the beginning of the pattern in each one of sequences (this corresponds to the vector x in the algorithm). Clearly this is a Markov chain. On the other hand, to give the beginning of the pattern is equivalent to give the pattern himself.

In this section our aim is to study this chain in order to detect the convergence properties of the algorithm.

2.3.1 The chain and its state space

Naturally, we are led to consider a sequence of random variables having vector values $(\hat{X}_n)_{n\geq 0}$ which describe the position of the beginning of the pattern on the nth iteration of the algorithm. Each component i of the vector $\hat{X}_n$ represents the position of the beginning of the pattern on the sequence i. The corresponding space state $\hat{E}$ is given by :

$$\hat{E} = \{(x_1, \ldots, x_N);\ x_i \in \{1, \ldots, L - W + 1\}\}, \tag{2.3.1}$$

where N notes the number of sequences, L the length of each sequence and x_i is the position of the beginning of the pattern on the sequence number i. It is thus obvious that :

$$\#\hat{E} = (L - W + 1)^N.$$

2.3.2 The transitions of the chain $\hat{X}$

The initial law of this chain is simple. Indeed, the position of the beginning of the pattern is chosen uniformly on each one of sequences and this is done independently on the set of sequences. So this initial law is :

$$\mathcal{L}(\hat{X}_0) = (U_1, U_2, \ldots, U_N) \tag{2.3.2}$$

where U_i are uniform random variables on $\{1, \ldots L - W + 1\}$.

It is clear that the only possible transitions are those for which only one component is modified, because in the algorithm, we shift the pattern only on one sequence. If we pursue exactly the algorithm, the transition probability p is given by the formula :

$$\begin{aligned} &p((x_1, x_2, \ldots, x_N), (y_1, y_2, \ldots, y_N)) \\ &\quad = \mathbb{P}\left(\hat{X}_{p+1} = (y_1, \ldots, y_N) \mid \hat{X}_p = (x_1, \ldots, x_N)\right) \end{aligned}$$

$$= \frac{1}{N}\left(\frac{V_{y_j}^{(j)}}{\sum_{r=1}^{L-W+1} V_r^{(j)}} \prod_{\{l=1,l\neq j\}}^{N} \mathbb{1}_{\{x_l=y_l\}}\right) \tag{2.3.3}$$

with

$$V_y^{(j)} = \frac{Q^{(j)}(1,A(j,y))\ Q^{(j)}(2,A(j,y+1))\dots Q^{(j)}(W,A(j,y+W-1))}{p(A(j,y))p(A(j,y+1))\dots p(A(j,y+W-1))}, \tag{2.3.4}$$

where for $l = 1,\dots,W$ and $k \in \mathcal{A}$

$$Q^{(j)}(l,k) = \frac{C_{l,k}^{(j)} + p(k)\sqrt{N}}{N-1+\sqrt{N}}, \tag{2.3.5}$$

$$C_{l,k}^{(j)} = \sum_{r=1,r\neq j}^{N} \mathbb{1}_{\{A(r,x(r)+l-1)=k\}}. \tag{2.3.6}$$

Remark 28 The term $1/N$ comes from the choice of the sequence. Indeed, not only one component changes on each stage, but also the choice of this component is done uniformly on the set of N sequences, in an independent way. If we call Y the random variable corresponding to the choice of the sequence, we have

$$\begin{aligned}
&\mathbb{P}\left(\hat{X}_{p+1} = (x_1,\dots,x_{j-1},y_j,x_{j+1},\dots,x_N) \mid \hat{X}_p = (x_1,\dots,x_N)\right),\\
&= \mathbb{P}(Y=j)*\\
&\qquad \mathbb{P}\left(\hat{X}_{p+1} = (x_1,\dots,x_{j-1},y_j,x_{j+1},\dots,x_N) \mid \hat{X}_p = (x_1,\dots,x_N) \cap Y = j\right),\\
&= \frac{1}{N}\left(\frac{V_{y_j}^{(j)}}{\sum_{r=1}^{L-W+1} V_r^{(j)}}\right).
\end{aligned}$$

We can check without difficulty that we obtain a stochastic matrix. The sum of all the coefficients on each row is :

$$\sum_{i=1}^{N} \frac{1}{N} \sum_{y_i=1}^{L-W+1} \frac{V_{y_i}^{(i)}}{\sum_{y_i=1}^{L-W+1} V_r^{(i)}} = 1.$$

Let us also note that the probability to rest in the same state $(x_1,\dots,x_n)$ is :

$$\frac{1}{N}\sum_{i=1}^{N} \frac{V_{x_i}^{(i)}}{\sum_{r=1}^{L-W+1} V_r^{(i)}}.$$

Indeed the probability to rest in the same state is the sum for all $i \in \{1,\dots N\}$ to rest in the same state conditionally on choosing to modify the sequence i.

We are now able to state the following important proposition :

Proposition 29 *The chain $\hat{X}$ has a stationary distribution.*

Démonstration In order to obtain the existence of a stationary distribution, we shall prove that the chain is homogeneous and recurrent.

As we can write the transition probabilities independently on the stage we are, the chain is clearly homogeneous

Furthermore $\hat{X}$ is irreducible.The probability to pass from a state

e_1 to the state e_2, when only one component changes between these tow states, is nonnegative, as

$$Q^{(j)}(l,k) = \frac{C_{l,k}^{(j)} + p(k)\sqrt{N}}{N - 1 + \sqrt{N}}.$$

Without restriction of generality we can assume that $p(k) > 0$. Otherwise, it means that the letter k does not appear in the proposed sequences so we can erase it from the state space. Then, for all k, j, l we have $Q^{(j)}(l,k) > 0$.

Thus, it is always possible to pass, in a finite number of steps, from some state e_1 to a state e_2 : it suffices to change the components one by one.

The chain is so irreducible and the state space is finite, so $\hat{X}$ is recurrent.

We conclude thus that $\hat{X}$ has a stationary distribution.

□

In order to obtain this stationary distribution, we are led to write the transition matrix Π of the chain $(\hat{X}_n)_{n\geq 0}$. $\hat{X}$ has vectorial values, so we will define an equivalent chain X with scalar values, for which one we re-ordinate the states.

2.3.3 The chain X

It is more convenient to work on a state space with no-vectorial components. We associate in a natural manner, to $\hat{E}$ a space E in which we re-ordinate the states according to the lexicographical order :

$$E = \left\{0, 1, \ldots, (L - W + 1)^N - 1\right\}.$$

It is obvious that $\#E = (L - W + 1)^N$. Let us express the bijection between $\hat{E}$ and E.
Let ϕ the application defined by

$$\phi \ : \ \hat{E} \longrightarrow E, \tag{2.3.7}$$

$$\phi((x_1, \ldots, x_N)) = \sum_{i=1}^{N} (x_i - 1)(L - W + 1)^{N-i}. \tag{2.3.8}$$

Inversely, in order to pass from E to $\hat{E}$, it suffices to determine by recurrence the different components of the vector $(x_1, \ldots, x_N)$. Indeed,

given an element k from E, $x_N - 1$ is the rest of the Euclidean division of k by $L - W + 1$.
$x_N, \ldots, x_{N-j}$ given, $x_{N-j-1} - 1$ is the rest of the Euclidean division of

$$k - \big((x_N - 1) + (x_{N-1} - 1)(L - W + 1) + \ldots + (x_{N-j} - 1)(L - W + 1)^j\big) \tag{2.3.9}$$

by $(L - W + 1)^{j+2}$.
Another manner to obtain the component x_j is :

$$x_j - 1 = \left[\frac{k}{(L - W + 1)^{N-j}}\right] - (L - W + 1)\left[\frac{k}{(L - W + 1)^{N-j+1}}\right], \tag{2.3.10}$$

where [.] denotes the integer part.

We define the chain X by

$$X = \phi(\hat{X}), \tag{2.3.11}$$

where ϕ is the application defined on (2.3.7). Then X takes it's values on E.

2.3.4 The transition matrix of the chain X

Initially we will search to identify the impossible transitions. We saw in the section 2.3.2 that the only possible transitions for $\hat{X}$ were those where at most one component has been changed.

Let us consider given a state $(x_1, \ldots, x_N)$ of the chain $\hat{X}$. Let

$$k = \sum_{i=1}^{N} (x_i - 1)(L - W + 1)^{N-i}$$

the corresponding state of the chain X. Then :

$$(x_1 - 1)(L - W + 1)^{N-1} \leq k \leq x_1(L - W + 1)^{N-1} - 1. \tag{2.3.12}$$

Let $(y_1, \ldots, y_N) \in \hat{E}$ and $k' = \phi((y_1, \ldots, y_N))$. **1)** Suppose that $x_1 \neq y_1$. Then

$$k \notin [\![(x_1 - 1)(L - W + 1)^{N-1},\ x_1(L - W + 1)^{N-1} - 1]\!].$$

The unique possibility that the transition from k to k' takes place with a nonnegative probability is that $(x_2, \ldots, x_n) = (y_2, \ldots, y_n)$, that means the rest of the Euclidean division of k by $(L - W + 1)^{N-1}$ equals the rest of the Euclidean division of k' by $(L - W + 1)^{N-1}$. **2)** If $x_1 = y_1$ then we are led to consider the possible transitions from $(x_2, \ldots, x_n)$ to $(y_1, \ldots y_n)$.

We can remake the previous reasoning for the transitions of $(x_2, \ldots, x_n)$ to $(y_1, \ldots y_n)$.

Description of the transition matrix By convention we note with D_l a diagonal matrix of type $l \times l$ having all the elements on the diagonal nonnegative.

The transition matrix of the chain has a quite simple structure. It can be decomposed in blocks in the following way. **1)** We consider the blocks of size $(L-W+1)^{N-1} \times (L-W+1)^{N-1}$. They are $(L-W+1)^2$ such blocks. Among these $(L-W+1)^2$ blocks there are $L-W+1$ matrices of size $(L-W+1)^{N-1} \times (L-W+1)^{N-1}$ on the "diagonal". The other blocks will be of type $D_{(L-W+1)^{N-1}}$.

2) We continue on this manner by considering blocks on the "diagonal" of the previous step. These ones will split in $(L-W+1)^2$ blocks of size $(L-W+1)^{N-2} \times (L-W+1)^{N-2}$ and we find the same structure as in **1)**. For the non-diagonal blocks we have matrices of type $D_{(L-W+1)^{N-2}}$.

3) This decomposition follows up on this manner until obtaining blocks of size $(L-W+1) \times (L-W+1)$. The blocks situated on the "diagonal" are matrices having all elements nonnegative and the other ones are blocks of type D_{L-W+1}.

Let us made some remarks concerning this matrix.

Remark 30 First of all it is interesting to remark that the structure of the matrix is closely related to the description (2.3.9) of the chain X by using the Euclidean division with respect to the powers of $L-W+1$.

Remark 31 When $x_1 \neq y_1$ this means that we are in a block of size $(L-W+1)^{N-1}$ which is not on the diagonal. Indeed we are on a row k such that :

$$(x_1-1)(L-W+1)^{N-1} \leq k < (x_1-1)(L-W+1)^{N-1}$$

and on a column k' with

$$(y_1-1)(L-W+1)^{N-1} \leq k < (y_1-1)(L-W+1)^{N-1}.$$

So, by using previous results, the only possible transitions are those which keep the components $(x_2, \ldots, x_N)$ unchanged. This explains the diagonal structure of blocks of type $D_{(L-W+1)^{N-1}}$.

In order to simplify this idea, let us describe the steps in the following diagram, in which we used the notation $D^{(l)}$ for $D_{(L-W+1)^l}$.

For illustrating clearly the structure of the matrix let us consider the precise following precise situation (figure 2).

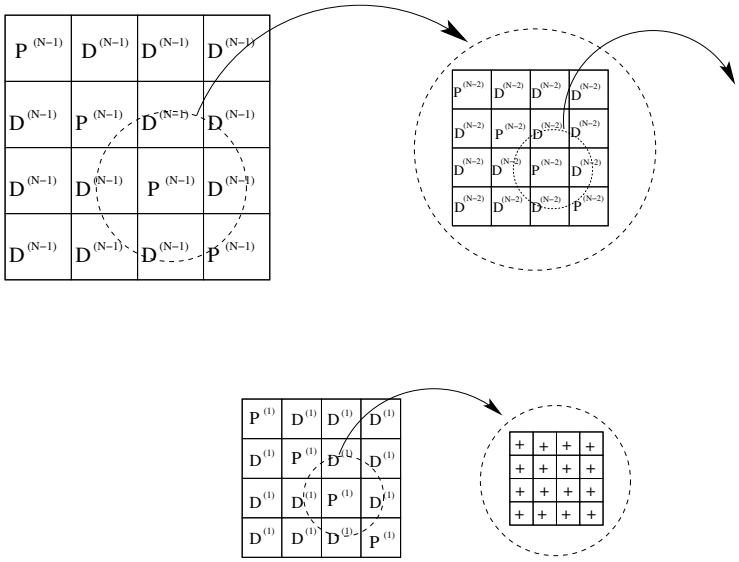

Figure 2.1: Diagram

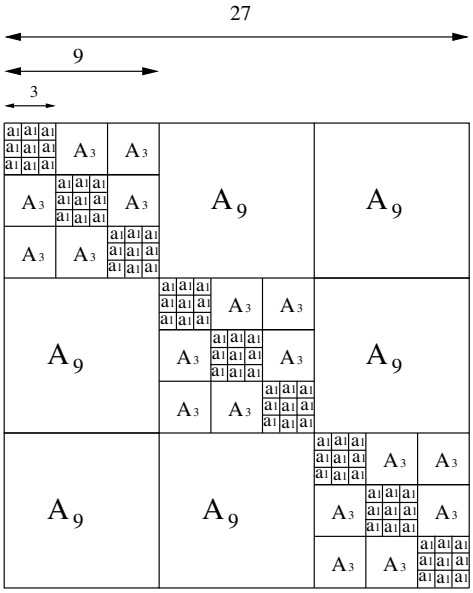

Figure 2.2: An example : $L - W + 1 = 3$, $N = 3$.

2.4 appendix

2.4.1 Algorithm

```
 /* Initialisation of the vector X which contains the beginning
of the patterns in each sequence.
we initialise by random positions */
 For i=0; i<=W-1; i++
    {pattern[i]=0;
    }
  /* size of the longest sequence */
  srand((unsigned)time(NULL));

  For (i=0; i<=N; i++)                  | we initialise the
    {                                   | beginning of the pattern
      u=random number on 0, L-W+1       | on each sequence
      X[i]=u;
    }

  /* we evaluate the probabilities p[i] of the appearance of each letter */

  For (i=0; i<=3; i++)   /* initialisation */
    {
      p[i]=0;
    }

  For (i=0; i<=N-1; i++)
    { For 0; j<=L-1; j++)
      {
        p[A[i][j]-1]=p[A[i][j]-1]+1;
        }
      }
    }
  For (i=0; i<=3; i++)   /* initialisation */
    {
      p[i]=p[i]/NL;
    }

  /* We iterate the algorithm in order to find the pattern */
  For (iter=1; iter<=nbre_iter; iter++)
    {For0; i<=3; i++)
      { For; j<=W-1; j++)
        { Q[i][j]=(double)0;
        }
```

```
  }
u=random number between 1 and N;
choice=u;

/* Part description of the pattern, we fill in Q */
For;i<=W-1;i++)
  { For;k<=3; k++)
    { For0; j<=N-1; j++)
      { If (j==choice) then
        {  If[j][X[j]+i]==k+1)
            { Q[k][i]= Q[k][i]+1;
            }
        }
      }
    Q[k][i]=(Q[k][i]+(p[k]*exp(0.5*log(N))))/(Nbre_seq-1+exp(0.5*log(N)));
    }
  }

/* Step: searh the pattern in the choosed sequence */
For(i=0;i<=L-1;i++)
  {Pos[i]=(double)1;
  }
For(i=0;i<=L-W;i++)
  { If(A[choice][i])
    {For(j=0;j<=W-1;j++)
      {If(A[choice][i+j])
        {
        Pos[i]=Pos[i]*(Q[A[choice][i+j]-1][j])/(p[A[choice][i+j]-1]);
        }
      and
        {
          Pos[i]=0;
          break;}
      }
    }
  else
    {
    break;}
  }

low=Pos;
For(i=0;i<=max_legth-1;i++)
  { if(i)
      Pos[i]=Pos[i]+Pos[i-1];
```

```
      }
    lower_bound=0;
    upper_bound=real_length-W;
    u=number between 1 and L;
    until (!(lower_bound==upper_bound-1))
      {If(u<=Pos[(int)((double)(upper_bound+lower_bound)/2)])
        {upper_bound= (int)((double)(upper_bound+lower_bound)/2);
        }
      Else
        {lower_bound = (int)((double)(upper_bound+lower_bound)/2);
        }
      }
    X[choice]=upper_bound;
    Si(low[[choice]]>likelihood)
      {likelihood=loi[upper_bound];
      For(j=0;j<=3;j++)
        {pattern[j]=A[choice][X[choice]+j];        }
      }
    }
}
```

Bibliography

[LSB+93] C. E. Lawrence, S.Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wooton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

Conclusion générale