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 Appliquées**

par **Marie-Pierre ETIENNE**

Le score local, Un outil pour l'analyse de séquences biologiques.

soutenue le 20 décembre 2002

devant le jury composé de Messieurs les Professeurs :

Président :	Aimé LACHAL	PRAG à l'INSA de Lyon
Rapporteurs :	Jean-Marc AZAIS	Professeur à l'Université Paul Sabatier Toulouse III
	Bernard PRUM	Professeur à l'Université d'Évry Val d'Essone
Examinateurs :	Dominique CELLIER	Maître de Conférence à l'Université de Rouen
	Bernard ROYNETTE	Professeur à l'Université Henri Poincaré Nancy I
	Pierre VALLOIS	Professeur à l'Université Henri Poincaré Nancy I

Institut Élie Cartan Nancy

Remerciements

Je tiens tout d'abord à remercier Pierre Vallois pour m'avoir encadrée et guidée durant les trois années de ma thèse. Ses conseils précieux et sa disponibilité constante m'ont été d'une grande aide.

J'ai eu l'occasion de travailler avec Jean-Jacques Daudin à de nombreuses reprises et cette collaboration s'est toujours montrée très enrichissante. Je voudrais le remercier pour m'avoir initiée au travail sur les séquences biologiques.

Un grand merci également à l'équipe de probabilités de Nancy. L'ambiance chaleureuse qui y règne permet de nombreux échanges et les pauses café se sont souvent révélées très enrichissantes d'un point de vue scientifique. Je tiens à remercier tout particulièrement Régine Marchand, Madalina Deaconu et Nicolas Fournier qui m'ont toujours manifesté le plus grand soutien et qui ont montré une patience sans faille à mon égard.

Je suis actuellement dans l'équipe Statistique et Génome et je tiens à en remercier tous les membres, non seulement pour l'accueil chaleureux qu'ils m'ont réservé dès mon arrivée mais également pour l'aide qu'ils m'ont apportée en me faisant partager leurs connaissances en bioinformatique. Je remercie particulièrement le directeur de cette équipe, Bernard Prum, qui a accepté de rapporter ma thèse et dont les connaissances m'ont été d"une grande aide dans la phase de rédaction.

Je souhaite aussi remercier Jean-Marc Azaïs qui a rapporté ma thèse, ainsi que tous les membres de mon jury Dominique Cellier, Aymé Lachal et Bernard Roynette qui m'ont permis de donner la forme finale à ce document.

Enfin mes derniers remerciements vont naturellement à ma famille et particulièrement à mes parents sans qui je n'aurais jamais pu entreprendre ce travail de recherche et dont la présence et le soutien constant, notamment durant les dernières semaines avant la soutenance, ont représenté plus que je ne saurais dire.

Table des matières

Introduction générale

L'ADN d'un être vivant contient toutes les informations génétiques nécessaires à son développement. C'est une molécule très longue composée de quatre bases élémentaires. La succession de ces bases contient de l'information sous forme codée. Il existe d'autres types de séquences biologiques, chacune ayant ses propres caractéristiques et ses propres fonctions (cf. chapitre 1, partie I, page 5.). Néanmoins, on peut toutes les considérercomme des chaînes de caractères sur un alphabet fini. Les nombreuses correspondances entre ces différentes biomolécules nous incitent à les étudier conjointement.

Actuellement, on dispose des séquences de plusieurs génomes complets, les bases de données de biomolécules s'enrichissent chaque jour et un traitement manuel de ces données est devenu inenvisageable. Aussi est-il nécessaire de mettre en place des outils informatiques et mathématiques permettant de systématiser le traitement des séquences et de débroussailler le champ de recherche pour le biologiste. L'un de ces outils est le score local, il est utilisé largement dans la comparaison de séquences et dans la recherche de propriétés physico-chimiques des protéines.

Le travail présenté ici porte sur l'étude du score local en tant qu'objet mathématique. Ce mémoire est composé de deux parties distinctes.

La première a pour but de définir les notions de base de biologie nécessaires à la compréhension du problème. De nombreux exemples illustrent chaque notion et aucune connaissance préalable en biologie n'est demandée. Cette partie met l'accent sur les difficultés rencontrées lors de l'analyse de séquences biologiques et la multitude de façons d'aborder le problème, ne serait-ce que concernant l'analyse de la similarité de séquences biologiques. Dans ce cadre l'utilisation du score local est devenue classique.

La seconde partie est dédiée à l'étude du score local en tant que variable aléatoire. Cette analyse va permettre de donner un degré de significativité statistique aux scores obtenus dans la pratique et de répondre à la question "le score obtenu est-il surprenant ou non ?" . On verra dans le chapitre 4 que le score local peut-être approché par des fonctionnelles browniennes. Il existe d'autres approximations pour le score local ; le travail présenté chapitre 5 présente les comparaisons numériques de ces différentes approximations et donne quelques clés pour le choix de l'approximation la mieux appropriée.

Les constatations numériques conduisent naturellement à évaluer la vitesse de convergence du score local dans le cas où les variables aléatoires sous-jacentes sont centrées. Plus généralement, nous étudierons, dans le chapitre 6, la vitesse de convergence du maximum d'une marche aléatoire vers sa limite : le maximum du mouvement brownien sur $[0;1]$.Les mêmes techniques nous permettront de déterminer la vitesse de convergence du score local.

Pour faciliter l'accès à l'information dans ce mémoire, j'ai choisi de mettre une table des matières générale mais également de rappeler au début de chaque chapitre la table des matières propre à ce chapitre. Dans la même optique, chaque chapitre de la partie II possède sa propre bibliographie afin de classer les références selon leur utilisation.

Première partie

Biologie et probabilités

Chapitre 1

Quelques éléments de biochimie

Contents

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].

Après avoir défini les trois types de séquences biologiques que l'on sera susceptible de rencontrer, nous verrons dans le paragraphe 1.2 les relations étroites qui les lient les unes aux autres. Nous aborderons ensuite les différents types d'évolution que peuvent subir les séquences. Enfin, la dernière partie de ce chapitre donnera quelques exemples de séquences, elle permettra également de donner des idées sur la taille des séquences considérées.

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, dont la structure fut 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. On les classe en deux catégories : les bases puriques (adénine et guanine) et les bases pyrimidiques (cytosine et thymine). Cette répartition correspond à des propriétés chimiques de ces bases.

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 orientés ; les biologistes appellent 5′ le côté initial et 3′ le côté terminal. 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 deux brins d'ADN sont orientés selon des directions opposées : c'est l'anti-parallélisme. On peut résumer ces propriétés sur un schéma 1.1 : On peut alors représenter un morceau d'ADN comme

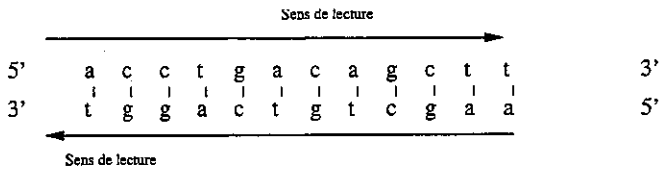

FIG. 1.1 – Complémentarité et anti-parallélisme de l'ADN.

dans la figure 1.2.

1.1.2 l'ARN : Acide Ribonucléique

L'ARN est assez semblable à l'ADN mais il ne comporte qu'un unique brin.Cela lui donne une plus grande souplesse de structure Cette molécule et elle a tendance à se replier sur elle-même de diverses manières. L'ARN se décrit également à partir d'un alphabet à quatre lettres {a, u, g, c}, où l'uracile (u) remplace la thymine. Au contraire de l'ADN, il y a différents

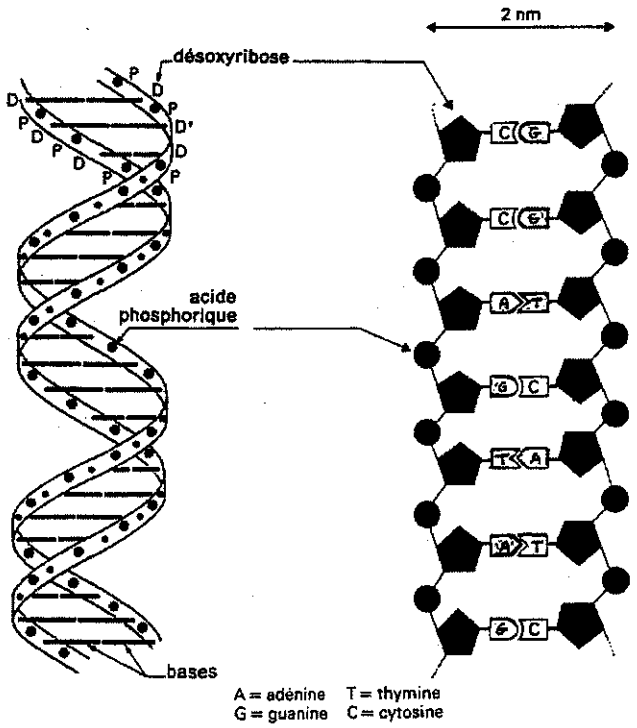

FIG. 1.2 – La double hélice et la complémentarité de l'ADN.

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 l'information génétique contenu dans l'ADN en utilisant le code génétique.

Acide Aminé	Abr		Acide Aminé	Abr	
alanine	Ala	A	méthionine	Met	M
cystéine	Cys	C	aspargine	Asn	N
acide aspartique	Asp	D	proline	Pro	P
acide glutaminique	Glu	E	glutamine	Gln	Q
phénylalanine	Phe	F	arginine	Arg	R
glycine	Gly	G	sérine	Ser	S
histine	His	H	thréonine	Thr	T
isoleucine	Ile	I	valine	Val	V
lysine	Lys	K	tryptophane	Trp	W
leucine	Leu	L	tyrosine	Tyr	Y

On peut ajouter deux lettres à cet alphabet. En effet l'asparagine (N) et l'acide aspartique (D) sont très similaires, on les regroupe parfois sous la lettre B (Asx). Il en va de même pour la glutamine (Q) et l' acide glutamique regroupés sous la lettre Z (Glx).

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 associés à leur étude.

1.2 De l'ADN aux protéines

La synthèse de protéines s'effectue aux travers de deux 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 3% 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 (ARNm) porte le code génétique.

1.2.2 Traduction : de l'ARNm aux protéines

Ce mécanisme met en jeu un complexe appelé ribosome qui se fixe sur l'ARNm pour traduire à l'aide d'un autre type d'ARN, l'ARN de transfert (ARNt) 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

On peut remarquer qu'un acide aminé peut être codé de plus d'une façon : par exemple la sérine (Ser ou S) peut être codée par tca, tcc, tcg, tct, agc ou agt. On dit que le code génétique est dégénéré. On sait que selon les espèces, certains codons sont utilisés de manière préférentielle mais les raisons du choix d'un codage au détriment d'un autre sont encore mal connues.

1.3 Évolution des séquences

Le matériel génétique évolue au cours du temps de plusieurs manières. Des erreurs de transcription ou de traduction peuvent se produire : elles sont parfois 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 l'information génétique qu'il contient permet encore la synthèse de protéines viables, alors cet ADN est conservé et crée la variabilité génétique. Les altérations de l'ADN sont multiples, on s'intéresse essentiellement à deux d'entre elles : les substitutions et les insertions délétions que l'on regroupe sous le terme indels.

1.3.1 Substitutions

On appelle **substitution** le fait de remplacer un nucléotide (ou un acide aminé) 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

On distingue habituellement deux types de substitutions :
- la substitution d'une base purique par une base purique (a / g) ou d'une base pyrimidique par une base pyrimidique (c / t) qui constitue une transition..
- les remplacements d'une base purique par une base pyrimidique et réciproquement (a,g / c,t) qui sont appelés des transversions.

Ces deux types de substitutions ont en effet des conséquences assez différentes puisqu'une transition sur la troisième base d'un codon est sans conséquence dans 90% des cas (à cause de la dégénérescence du code génétique) contre seulement 30% dans le cas d'une transversion.

1.3.2 Délétions et insertions

Un autre type d'altérations de l'information génétique est la suppression d'un ou plusieurs nucléotides. C'est-à-dire qu'on passe de la séquence A à la séquence B en supprimant un nucléotide.

Exemple 1.1 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 mutation sous le nom d'insertions-délétions ou encore **indels**.

Il existe d'autres types de transformations de l'ADN. Parmi les plus importantes il convient de citer
- les inversions, pour lesquelles un morceau d'ADN provenant du brin complémentaire vient s'insérer dans le brin étudié, ce morceau de séquence se lit donc dans le sens contraire.

organisme	taille	description
Haemophilus influenzae	1.8 Mb	bacille infectieux
Mycoplasma genitalium	0.6 Mb	parasites des voies génitales
Human immunodeficiency virus 1	0.01 Mb	virus HIV
Hepatitis B virus	0.03Mb	virus de l'hépathite B
Homo sapiens	3100 Mb	l'homme
Escherichia coli	4.6Mb	bacille modèle

TAB. 1.3 – Taille de quelques génomes exprimée en millions de bases (Mb)

- les recombinaisons pour lesquelles deux chromosomes échangent une partie de leur information.
- le réarrangement génétique qui consiste à réunir plusieurs morceaux d'ADN initialement non contigus.
- ...

Ces éléments biologiques sont essentiels à la construction d'un être vivant. Ils 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.

Lorsqu'on cherche à étudier des séquences, on étudie en fait une chaîne de caractères dont les lettres sont dans un alphabet à quatre ou vingt lettres. Pour fixer les idées, il est important de bien comprendre ce que sont les séquences biologiques et d'avoir une idée de leur taille. Celle-ci est très différente selon que l'on s'intéresse aux génomes entiers ou aux protéines. Ce qui suit présente quelques données relatives aux séquences et à leurs tailles.

1.4 Des exemples de séquences

1.4.1 La taille de certains génomes

Le tableau 1.3 présente la taille des génomes de certains organismes. Il faut remarquer que la taille d'un génome d'un être vivant est très variable ; elle n'est que relativement liée à la complexité de l'être vivant qui lui est associé.

1.4.2 Un exemple de génome : Le virus de l'hépatite B

Nous avons choisi de présenter dans son intégralité le plus petit des génomes cités ci-dessus, celui du virus de l'hépatite B. Il est au format FASTA, un format classique pour la notation des séquences biologiques. Il est difficile d'extraire une information de la lecture lettre à lettre de ce génome, mais il est important de se faire une idée visuelle de ce que sont les séquences considérées.

```
>gi|21326584|ref|NC_003977.1| Hepatitis B virus, complete genome
CTCCACAACATTCCACCAAGCTCTGCTAGATCCCAGAGTGAGGGGCCTATATTTTCCTGCTGGTGGCTCC
AGTTCCGGAACAGTAAACCCTGTTCCGACTACTGCCTCACCCATATCGTCAATCTTCTCGAGGACTGGGG
ACCCTGCACCGAACATGGAGAGCACAACATCAGGATTCCTAGGACCCCTGCTCGTGTTACAGGCGGGGTT
TTTCTTGTTGACAAGAATCCTCACAATACCACAGAGTCTAGACTCGTGGTGGACTTCTCTCAATTTTCTA
GGGGGAGCACCCACGTGTCCTGGCCAAAATTCGCAGTCCCCAACCTCCAATCACTCACCAACCTCTTGTC
CTCCAACTTGTCCTGGCTATCGCTGGATGTGTCTGCGGCGTTTTATCATATTCCTCTTCATCCTGCTGCT
ATGCCTCATCTTCTTGTTGGTTCTTCTGGACTACCAAGGTATGTTGCCCGTTTGTCCTCTACTTCCAGGA
ACATCAACTACCAGCACGGGACCATGCAGAACCTGCACGATTCCTGCTCAAGGAACCTCTATGTTTCCCT
CTTGTTGCTGTACAAAACCTTCGGACGGAAACTGCACTTGTATTCCCATCCCATCATCCTGGGCTTTCGC
AAGATTCCTATGGGAGTGGGCCTCAGTCCGTTTCTCCTGGCTCAGTTTACTAGTGCCATTTGTTCAGTGG
TTCGTAGGGCTTTCCCCCACTGTTTGGCTTTCAGCTATATGGATGATGTGGTATTGGGGGCCAAGTCTGT
ACAACATCTTGAGTCCCTTTTTACCTCTATTACCAATTTTCTTTTGTCTTTGGGTATACATTTGAACCCT
AATAAAACCAAACGTTGGGGCTACTCCCTTAACTTCATGGGATATGTAATTGGAAGTTGGGGTACTTTAC
CGCAGGAACATATTGTACAAAAACTCAAGCAATGTTTTCGAAAATTGCCTGTAAATAGACCTATTGATTG
GAAAGTATGTCAAAGAATTGTGGGTCTTTTGGGCTTTGCTGCCCCTTTTACACAATGTGGCTATCCTGCC
TTGATGCCTTTATATGCATGTATACAATCTAAGCAGGCTTTCACTTTCTCGCCAACTTACAAGGCCTTTC
TGTGTAAACAATATCTAAACCTTTACCCCGTTGCCCGGCAACGGTCAGGTCTCTGCCAAGTGTTTGCTGA
CGCAACCCCCACGGGTTGGGGCTTGGCCATAGGCCATCGGCGCATGCGTGGAACCTTTGTGGCTCCTCTG
CCGATCCATACTGCGGAACTCCTAGCAGCTTGTTTTGCTCGCAGCCGGTCTGGAGCGAAACTTATCGGAA
CCGACAACTCAGTTGTCCTCTCTCGGAAATACACCTCCTTTCCATGGCTGCTAGGCTGTGCTGCCAACTG
GATCCTGCGCGGGACGTCCTTTGTCTACGTCCCGTCGGCGCTGAATCCCGCGGACGACCCGTCTCGGGGC
CGTTTGGGCCTCTACCGTCCCCTTCTTCATCTGCCGTTCCGGCCGACCACGGGGCGCACCTCTCTTTACG
CGGTCTCCCCGTCTGTGCCTTCTCATCTGCCGGACCGTGTGCACTTCGCTTCACCTCTGCACGTAGCATG
GAGACCACCGTGAACGCCCACCAGGTCTTGCCCAAGGTCTTACACAAGAGGACTCTTGGACTCTCAGCAA
TGTCAACGACCGACCTTGAGGCATACTTCAAAGACTGTTTGTTTAAAGACTGGGAGGAGTTGGGGGAGGA
GATTAGGTTAAAGGTCTTTGTACTAGGAGGCTGTAGGCATAAATTGGTCTGTTCACCAGCACCATGCAAC
TTTTTCCCCTCTGCCTAATCATCTCATGTTCATGTCCTACTGTTCAAGCCTCCAAGCTGTGCCTTGGGTG
GCTTTGGGGCATGGACATTGACCCGTATAAAGAATTTGGAGCTTCTGTGGAGTTACTCTCTTTTTTGCCT
TCTGACTTCTTTCCTTCTATTCGAGATCTCCTCGACACCGCCTCTGCTCTGTATCGGGAGGCCTTAGAGT
CTCCGGAACATTGTTCACCTCACCATACAGCACTCAGGCAAGCTATTCTGTGTTGGGGTGAGTTGATGAA
TCTGGCCACCTGGGTGGGAAGTAATTTGGAAGACCCAGCATCCAGGGAATTAGTAGTCAGCTATGTCAAT
GTTAATATGGGCCTAAAAATTAGACAACTATTGTGGTTTCACATTTCCTGCCTTACTTTTGGAAGAGAAA
CTGTCCTTGAGTATTTGGTGTCTTTTGGAGTGTGGATTCGCACTCCTCCCGCTTACAGACCACCAAATGC
CCCTATCTTATCAACACTTCCGGAAACTACTGTTGTTAGACGACGAGGCAGGTCCCCTAGAAGAAGAACT
CCCTCGCCTCGCAGACGAAGGTCTCAATCGCCGCGTCGCAGAAGATCTCAATCTCGGGAATCTCAATGTT
AGTATCCCTTGGACTCATAAGGTGGGAAACTTTACTGGGCTTTATTCTTCTACTGTACCTGTCTTTAATC
CTGATTGGAAAACTCCCTCCTTTCCTCACATTCATTTACAGGAGGACATTATTAATAGATGTCAACAATA
TGTGGGCCCTCTGACAGTTAATGAAAAAAGGAGATTAAAATTAATTATGCCTGCTAGGTTCTATCCTAAC
CTTACCAAATATTTGCCCTTGGACAAAGGCATTAAACCGTATTATCCTGAATATGCAGTTAATCATTACT
TCAAAACTAGGCATTATTTACATACTCTGTGGAAGGCTGGCATTCTATATAAGAGAGAAACTACACGCAG
CGCCTCATTTTGTGGGTCACCATATTCTTGGGAACAAGAGCTACAGCATGGGAGGTTGGTCTTCCAAACC
TCGACAAGGCATGGGGACGAATCTTTCTGTTCCCAATCCTCTGGGATTCTTTCCCGATCACCAGTTGGAC
CCTGCGTTCGGAGCCAACTCAAACAATCCAGATTGGGACTTCAACCCCAACAAGGATCACTGGCCAGAGG
CAAATCAGGTAGGAGCGGGAGCATTTGGTCCAGGGTTCACCCCACCACACGGAGGCCTTTTGGGGTGGAG
CCCTCAGGCTCAGGGCATATTGACAACACTGCCAGCAGCACCTCCTCCTGCCTCCACCAATCGGCAGTCA
GGAAGACAGCCTACTCCCATCTCTCCACCTCTAAGAGACAGTCATCCTCAGGCCATGCAGTGGAA
```

1.4.3 Une protéine : l'alpha hémoglobine

La taille des protéines n'est en rien comparable à la taille des génomes. Ce sont en général des séquences comportant quelques dizaines à quelques centaines d'acides aminés.

```
>gi|6230879|dbj|BAA86218.1| alpha hemoglobin A [Seriola quinqueradiata]
MSLSGKDKSVVKAFWDKMSPKSAEIGAEALGRMLTVYPQTKTYFSHWADVGPDSAQVKKHGATIMAAVGD
AVGKIDDLVGGLSALSELHAFKLRVDPANFRILAHNIILVTAMYFPTDFTPEIHVSVDKFLQNLALALAE
RYR
```

Il apparaît maintenant clairement que les tailles des différentes molécules biologiques étudiées sont très diverses.

Chapitre 2

Analyse de séquences

Contents

Comprendre le rôle de chacune des macromolécules citées dans le chapitre 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 un mécanisme connu.

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 fonction connue

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, les deux matrices les plus utilisées sont :

1. 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.

2. la matrice de Kimura

	a	c	g	t
a	γ	β	α	β
c	β	γ	β	α
g	α	β	γ	β
t	β	α	β	γ

où $\gamma > \alpha > \beta > 0$ et $\gamma = 1 - \alpha - 2\beta$.
Cette matrice pénalise davantage transversions que les transitions, ce qui permet de prendre en compte les conséquences très différentes de ces deux types de substitutions.

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 (indels de plusieurs bases) doivent être 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. Si l'approche est différente cette technique revient à calculer le score d'alignement de deux séquences à l'aide de la matrice identité. On peut donc se contenter d'utiliser la méthode du score pour déterminer la similarité de deux séquences.

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 trois opérations élémentaires.

Exemple 2.1 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 :

$$\begin{array}{lcccc} \mathrm{A}^* : & a_1^* & a_2^* & \ldots & a_L^* \\ \mathrm{B}^* : & b_1^* & b_2^* & \ldots & b_L^* \end{array}$$

Le nombre d'alignements possibles est asymptotiquement de l'ordre de $(1+\sqrt{2})^{m+n}$ où n et m désignent les longueurs respectives de A et B. En effet, lorsqu'on aligne une séquence **A** de taille n avec une séquence **B** de taille m, l'alignement peut se finir de trois manières :

$$\begin{array}{ccc} a_n & a_n & - \\ b_m & - & b_m \end{array}$$

Ainsi le nombre d'alignements possibles entre deux séquences n et m, noté $N(n,m)$ vérifie l'équation de récurrence suivante :

$$N(n,m) = N(n-1,m-1) + N(n-1,m) + N(n,m-1).$$

Une solution de cette équation se met sous la forme $N(n,m) = K^{n+m}$ où K doit vérifier

$$K^{n+m} = K^{n+m-2} + 2K^{n+m-1}.$$

Ce qui donne le résultat annoncé.

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 (a_i^*, b_i^*), on associe un score $s(a_i^*, b_i^*)$: le score élémentaire ; le score d'alignement noté $S(\mathbf{A}^*, \mathbf{B}^*)$ est la somme des scores de chaque paire.

Exemple 2.2 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 absence de pénalisation pour les brèches.

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 2.1.

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.

L'exemple qui suit présente une fonction de score qui permet de distinguer les transitions, transversions et les brèches.

Exemple 2.3 On définit une autre fonction de score s à valeurs dans $\mathbb{R}$, telle que :

$$s(a_i, b_j) = \begin{cases} 1 & \text{si } a_i = b_j, \\ 0 & \text{si la substitution } a_i \to b_j \text{ est une transition}, \\ -1 & \text{si la substitution } a_i \to b_j \text{ est une transversion}, \\ -2 & \text{dans le cas d'indel} \end{cases} .$$

Avec ce choix de fonction de score, on pénalise fortement les indels, tandis que les transitions sont neutres. On reprend l'alignement donné dans l'exemple 2.2.

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
	-2	1	-1	1	1	1	1	-2	1	1	1	1	1

Le score de cet alignement avec la nouvelle fonction de score considérée est donc 5.

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 optimal des séquences $\mathbf{A}$ et $\mathbf{B}$ sous la fonction de score s est par définition :

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

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

Étant 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 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 2.4 On reprend les séquences données dans l'exemple 2.1, un alignement local pourrait être le suivant :

$$\begin{array}{lccccc} \mathbf{A}: & \mathtt{t} & \mathtt{a} & \mathtt{g} & \mathtt{c} & \mathtt{t} \\ \mathbf{B}: & \mathtt{t} & \mathtt{a} & \mathtt{g} & \mathtt{c} & \mathtt{t} \end{array}$$

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. La complexité de cet algorithme reste la même que pour l'alignement global bien que le nombre de possibilités à explorer soit bien supérieur.

2.1.3 Alignement multiple

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 impliqué.

On peut également supposer qu'un ensemble de séquences est issu d'un même ancêtre commun et on aimerait mettre en évidence les régions de la séquence qui ont été bien conservées au cours de l'évolution. En effet lorsque l'on réussit à déterminer qu'une région est extraordinairement bien conservée, il est alors tentant de supposer qu'elle est à l'origine d'une fonction essentielle pour l'organisme considéré.

L'alignement multiple a donc pour but de mettre en évidence des régions très similaires. C'est la généralisation d'un alignement de deux séquences.

Les algorithmes d'alignement multiple

On distingue trois types d'approche pour l'alignement multiple.

1. La programmation dynamique

   Des algorithmes utilisant l'approche par programmation dynamique existent et fonctionnent lorsqu'on cherche à aligner un petit nombre de séquences entre elles. La complexité en temps de ce type d'algorithme pour aligner n séquences de longueur L est en $O((2L)^n)$. Ce qui signifie que le nombre de séquences à aligner est plus pénalisant en termes de temps d'éxécution que la longueur de celles-ci.

   Pour l'alignement local la situation est bien pire encore. Ainsi lorsque l'on cherche un détecter une petite portion commune dans plusieurs séquences, ce type d'algorithme n'est pas recommandé.

2. Approche heuristique

   On propose comme alternative l'algorithme heuristique suivant qui utilisée par de nombreux programmes d'alignement de séquences tels Clustal W [THG94] qui détermine une région très conservée mais n'affirme en aucun cas obtenir le meilleur motif commun à toutes les séquences. Après avoir déterminé les alignements deux à deux des séquences étudiées, il construit un arbre dans lequel les plus proches voisins sont les séquences les plus proches. Une fois l'arbre construit, on choisit les deux séquences les plus proches (qui auront le meilleur alignement) et on incorpore progressivement les autres séquences jusqu'à ajouter la séquence la plus distante. Ce programme est davantage sensible au nombre de séquences à aligner qu'à leur longueur. C'est-à-dire qu'il est

rapide lorsqu'on aligne un petit nombre de séquences longues mais plus lent lorsqu'on aligne un grand nombre de séquences plus courtes.

3. Un algorithme probabiliste
   Une autre approche est une approche probabiliste proposée par [LAB+93]. L'idée consiste à partir d'un motif quelconque et à l'affiner au fur et à mesure des itérations en utilisant des techniques de maximum de vraisemblance et d'échantillonage de Gibbs. L'avantage de cette méthode est que l'on travaille sur une seule séquence à la fois, ce qui induit des temps de calcul très raisonnables. Le principe de cet algorithme consiste à visiter aléatoirement différents alignements possibles mais n'assure pas d'obtenir le "meilleur" alignement local possible. Néanmoins après un certains nombres d'itérations, on repassera de plus en plus souvent par ce meilleur alignement et on devrait être capable de le reconnaître.

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.

Évidemment 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 (cf. 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 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. L'exemple qui suit (ex. 2.5) cherche à déterminer le segment le plus hydrophile d'une protéine.

Exemple 2.5 Hydrophobicité des protéines

Il est tout d'abord nécessaire de se donner une fonction de score. On va pour cela utiliser une échelle d'hydrophobicité (tableau 2.1).

R	K	D	E	N	Q	H	P	Y	W
-4.5	-3.9	-3.5	-3.5	-3.5	-3.5	-3.2	-1.6	-1.3	-0.9
S	T	G	A	M	C	F	L	V	I
-0.8	-0.7	-0.4	1.8	1.9	2.5	3.7	3.8	4.2	4.5

R, K, D, E sont donc hydrophiles, tandis que les acides aminés les plus hydrophobes sont M, C, F, L, V, I.

Le caractère hydrophobe ou hydrophile d'une protéine donne des informations indirectes sur sa fonction. En effet, on distingue deux type de protéines : les protéines membranaires et les protéines modulaires. La membrane est une paroi lipidique et donc hydrophobe. Les protéines membranaires présentent donc un ou plusieurs segments hydrophobes qui correspondent à la partie de la protéine en contact avec la membrane.

En ce qui concerne les protéines modulaires, le caractère hydrophobe d'un segment de protéine nous indique qu'il n'est pas en contact direct avec le milieu cellulaire, il est donc vraissemblablement enfermé à l'intérieur de la protéine.

On s'intéresse tout d'abord à l'insuline humaine dont la séquence est :

```
>gi|23821031|ref|NP_057217.1| insulin induced protein 2 [Homo sapiens]
MAEGETESPGPKKCGPYISSVTSQSVNLMIRGVVLFFIGVFLALVLNLLQIQRNVTLFPPDVIASIFSSA
WWVPPCCGTASAVIGLLYPCIDRHLGEPHKFKREWSSVMRCVAVFVGINHASAKVDFDNNIQLSLTLAAL
SIGLWWTFDRSRSGFGLGVGIAFLATVVTQLLVYNGVYQYTSPDFLYVRSWLPCIFFAGGITMGNIGRQL
AMYECKVIAENLIRNEEGKKYLLYRKAR
```

Le score global de cette séquence est 106.8 et le score local 152.5, le segment le plus hydrophobe correspond donc au segment qui réalise le score local c'est-à-dire le segment qui commence à la position 18 et se termine à la position 223.

Pour fixer les idées, il est utile de considérer maintenant une autre protéine : l'alpha hémoglobine humaine.

```
>gi|23813669|sp|Q9NZD4|AHSP_HUMAN Alpha-hemoglobin stabilizing protein
(Erythroid associated factor) (Erythroid differentiation related factor)
MALLKANKDLISAGLKEFSVLLNQQVFNDPLVSEEDMVTVVEDWMNFYINYYRQQVTGEPQERDKALQEL
RQELNTLANPFLAKYRDFLKSHELPSHPPPSS
```

On obtient un score global de -47.7 et un score local de 20. Le segment le plus hydrophobe s'étend de la position 10 à la position 22.

Lorsque l'on compare ces deux résultats, il semble que l'insuline est plus hydrophile que l'hémoglobine alpha, mais nous n'avons pour le moment aucune façon de donner un critère de significativité. C'est un des problèmes à l'origine de l'étude de la distribution du score local.

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 $q_{x_i} \times q_{y_j}$, où q_{x_i} est la probailité d'observer x_i dans une séquence. C'est ce qu'on appellera le modèle aléatoire (R). Ainsi la probabilité d'observer l'alignement est donnée par
$$\mathbb{P}(x, y \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}. \tag{2.3.1}$$
- 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 $q_{x_i}\pi_{x_iy_i}$ où q_{x_i} et la probabilité d'observer x_i est $\pi_{x_iy_i}$ est la probabilité de passer de x_i à y_i dans le modèle (M).

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 :

$$q_a\pi_{ab} = q_b\pi_{ba} = p_{ab} = p_{ba} \qquad \forall (a,b) \in \mathcal{A}^2. \tag{2.3.2}$$

Ainsi la probabilité d'observer l'alignement est donnée par

$$\mathbb{P}(x, y \mid M) = \prod_i q_{x_i}\pi_{x_iy_i} = \prod_i p_{x_iy_i}. \tag{2.3.3}$$

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 la rapport de vraisemblance.

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

En effet, on cherche à savoir ce qui est le plus probable ; si la rapport de vraisemblance est supérieur à un alors il est plus probable que l'on soit dans le cadre d'une évolution de séquences, tandis que si ce rapport est inférieur à un, on est conduit à penser que ces deux séquences n'ont rien à voir entre elles.

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 regroupe 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 PAMX (matrices de mutation de Dayhoff) : elles sont obtenues en calculant le logarithme des fréquences relatives de mutations de chaque acidé aminé. La PAM250 semble être la matrice la mieux adaptée pour distinguer les protéines apparentées (étude de [SD79]).

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

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

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	
4	-2	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2	-1	**A**
	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3	2	**B**
		9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2	-4	**C**
			6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3	2	**D**
				5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2	5	**E**
					6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3	-3	**F**
						6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3	-2	**G**
							8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2	0	**H**
								4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1	-3	**I**
									5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2	1	**K**
										4	2	-3	-3	-2	-2	-2	-1	1	-2	-1	-3	**L**
											5	-2	-2	0	-1	-1	-1	1	-1	-1	-2	**M**
												6	-2	0	0	1	0	-3	-4	-2	0	**N**
													7	-1	-2	-1	-1	-2	-4	-3	-1	**P**
														5	1	0	-1	-2	-2	-1	2	**Q**
															5	-1	-1	-2	-3	-2	0	**R**
																4	1	-2	-3	-2	0	**S**
																	5	0	-2	-2	-1	**T**
																		4	-3	-1	-2	**V**
																			11	2	-3	**W**
																				7	-2	**Y**
																					5	**Z**

TAB. 2.2 – La matrice BLOSUM62

Nous avons vu que pour améliorer un alignement de séquences, on devait 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 prsentés sont ceux qui sont utilisés en pratique dans les algorithmes d'alignement.

Chapitre 3

Significativité statistique

Contents

Comparer des séquences biologiques se ramène en fait à la comparaison de chaînes de caractères. Il est 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 pratique. Nous présenterons également deux autres approches plus théoriques : l'approche bayésienne et l'approche par les valeurs extrêmes. Enfin la dernière partie s'attachera spécifiquement 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. Cela revient à calculer un score global qui récompense par 1 une identité et ne pénalise ni les substitutions ni les indels. On normalise ensuite par la taille de la séquence. 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 vraisemblablement un ancêtre commun tandis que deux séquences nucléotidiques de 100 bases ayant plus de 50% d'identités n'ont pas forcément de lien biologique. cela provient du fait qu'une 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 pratique, 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 de réechantillonnage). 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.

On peut adopter la même démarche en préservant la composition de la séquence en mots de longueur 2 (modèle M1) ou de longueur 3 (modèle M2) ... Le choix du modèle dépend essentiellement des propriétés auxquelles on s'intéresse sur la séquence étudiée.

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	...	X_{n-1}	X_n	-	-
-	-	-	Y_1	Y_2	-	Y_3	...	...	...	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 fonctions strictement croissantes u et v de $\mathbb{N}$ dans $\{1,\dots,n\}$. Alors le score global de (X) et (Y) est donné par la formule :

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

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 $\mathbb{P}(M \mid X, Y)$ 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.

$$\mathbb{P}(M \mid X, Y) = \frac{\mathbb{P}(X, Y \mid M)\,\mathbb{P}(M)}{\mathbb{P}(X, Y)},$$

$$= \frac{\mathbb{P}(X,Y \mid M)\,\mathbb{P}(M)}{\mathbb{P}(X,Y \mid M)\,\mathbb{P}(M) + \mathbb{P}(X,Y \mid R)\,\mathbb{P}(R)},$$
$$= \frac{(\mathbb{P}(X,Y \mid M)\,\mathbb{P}(M))\,/\,(\mathbb{P}(X,Y \mid R)\,\mathbb{P}(R))}{(\mathbb{P}(X,Y \mid M)\,\mathbb{P}(M))\,/\,(\mathbb{P}(X,Y \mid R)\,\mathbb{P}(R)) + 1}.$$

On pose

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

où

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

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

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

avec

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

Etant donné un système de score, calculer $\mathbb{P}(X,Y \mid M)$ ou $\mathbb{P}(X,Y \mid R)$ ne présente pas de difficultés (cf. 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 élémentaires, 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 10^{-20} ou 10^{-40}), 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.

Le comportement du score local sur deux séquences (c'est-à-dire dans le cas d'un alignement) est mal connu, les résultats théoriques existants concernent essentiellement le score local sur une séquence et en pratique on utilise des variantes de ces résultats pour déterminer la significativité des scores d"alignement. 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 local.

3.5.1 Comportement asymptotique du score local

Nous énonçons ici le résultat de Karlin *et al* [DK92]. Ce résultat est un résultat clé puisqu'il est utilisé pour estimer le degré de significativité statistique des alignements obtenus dans le programme BLAST, qui est le programme d'alignement le plus utilisé actuellement.

Theorem 3.1 (Approximation de Karlin *et al.* [DK92]) Soit $(X_i)_{i\geq 1}$ une suite de variables indépendantes 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 d'espérances négatives. Dans le premier chapitre de la partie 2, nous montrerons que le score local dans le cas d'une espérance 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 (4.2.4) est convergente ([Bil68, p. 166]).

On a alors le résultat suivant pour le score local :

Theorem 3.2 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 exacte du score local

Daudin et Mercier [DM99] ont obtenu une formule exacte pour $\mathbb{P}(H_n < x)$ en utilisant une approche par les chaînes de Markov. Ainsi $\mathbb{P}(H_n < x)$ s'exprime à l'aide d'un vecteur P_n de longueur x donné par :

$$P_n = P_0 \Pi^n$$

où $P_0 = (1, 0, \ldots, 0)$ est de longueur x et π est la matrice de transition d'une chaîne de Markov à x états. Alors la distribution du score local est donnée par :

$$\mathbb{P}(H_n \geq x) = P_n(x). \tag{3.5.3}$$

En pratique cette formule n'est utilisable que pour x et n pas trop grands, sinon les temps de calculs recquis pour accéder à la distribution deviennent trop importants. Ces limitations sur ces deux données posent quelques problèmes. En effet, on s'intéresse souvent à de longues séquences (n grand), notamment lorsque l'on travaille sur des séquences nucléotidiques (cf. tableau 1.3) et aux événements rares, donc la queue de distribution. Il est donc important de disposer d'approximations pour n et x grands.

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 le calcul de 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.

[EG01] W.J. Ewens and G. R Grant. *Statistical methods in Bioinformatics.* Springer, 2001.

[KA90] S. Karlin and S. 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.

[LAB+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, 1993.

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

[Nue01] G. Nuel. *Grandes déviations et chaînes de Markov pour l'étude des occurrences de mots dans les séquences biologiques.* PhD thesis, Université d'Évry Val d'Esonne, juillet 2001.

[SD79] R. M. Schwarz and M. O Dayhoff. Matrices for detecting distant relationships. *Atlas of protein sequences*, pages 353–358, 1979.

[THG94] J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 1994.

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

Deuxième partie

Le score local

La partie qui suit est consacrée à l'étude du score local. Commençons par préciser le cadre de travail. Nous travaillerons sur une seule séquence. La figure 3.1 présente la modélisation de la séquence.

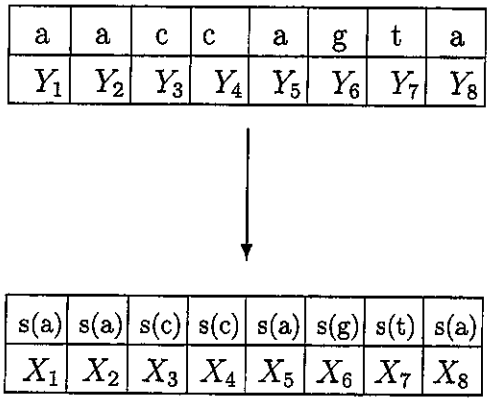

FIG. 3.1 – choix d'une statistique.(s fonction de score)

Ainsi dans toute la suite lorsque nous considérons le score local associé à la suite de variables aléatoirs $(X_n)_{n\geq 1}$, la fonction de score aura déjà été prise en compte. La loi des variables aléatoires considérées dépend donc à la fois de la modélisation initiale (en termes de variables aléatoires Y) et de la fonction de score considérée.

Cette deuxième partie est divisée en trois chapitres. Le premier présente un travail sur le comportement asymptotique du score local lorsque l'espérance des variables aléatoires considérées[1] est nulle ou "petite". On utilise des propriétés de convergence des marches aléatoires vers le mouvement brownien ainsi que des décompositions trajectorielles de celui-ci pour démontrer deux types de résultats.

- Dans le cas centré, on obtient une fonction de répartition limite pour le score local, qui est la fonction de répartition de la variable $\max_{0\leq u\leq 1}|B_u|$ où $(B_u, u \geq 0)$ désigne un mouvement brownien standard issu de 0.
- Dans le cas d'une espérance non nulle, on montre que la densité de la variable aléatoire vers laquelle converge le score local correctement normalisé vérifie une équation intégrale et on donne un équivalent pour la queue de distribution de cette variable aléatoire.

Le second chapitre compare les différentes approximations existantes pour la fonction de répartition du score, i.e. non seulement celles mises en évidence dans le chapitre quatre mais également l'approximation de Karlin

[1] Attention il s'agit bien de l'espérance des variables $(X_i)_{i\geq 1} = (s(Y_i))_{i\geq 1}$ et non de l'espérance des variables $(Y_i)_{i\geq 1}$ qui n'existe pas !

& al. Les constations numériques nous conduisent naturellement à étudier la vitesse de convergence du score local vers $\max_{0\leq u\leq 1}|B_u|$. Ce travail est présenté dans le dernier chapitre.

Chapitre 4

Comportement asymptotique du score local

Contents

Résumé :

Dans ce chapitre, nous allons présenter deux types de résultats de convergence pour le score local. Commençons par définir les objets que nous utiliserons.

On se donne tout d'abord une suite $(X_i)_{i\geq 1}$ de variables aléatoires. Soit $(S_n)_{n\geq 0}$ la marche aléatoire qui lui est associée :

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

On peut définir le score local de la suite $(X_n)_{n\geq 0}$ de la façon suivante :

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

On peut remarquer aisément que H_n s'écrit également

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

Cette écriture permet de donner l'intuition du résultat dans le cas centré en écrivant H_n comme une fonction de $(S_k)_{0\leq k\leq n}$. En effet, dans le cas où les variables aléatoires (X_i) considérées sont des variables indépendantes, centrées et de carré intégrable. $(S_n)_{n\geq 0}$ convenablement renormalisée, converge vers le mouvement brownien. On en déduit (cf théorème 4.1 page 49) :

$$\frac{H_n}{\sqrt{n}} \xrightarrow[n\to\infty]{(d)} \sigma \max_{0\leq u\leq 1}\left(B_u - \min_{0\leq v\leq u} B_v\right) \overset{(d)}{=} \sigma \max_{0\leq u\leq 1}|B_u|,$$

où σ désigne l"écart type de X_1.

Dans le cas non centré, on considère des variables aléatoires $(X_i^{(N)})_{i\geq 1}$ qui dépendent d'un paramètre N. On suppose que ces variables sont indépendantes, identiquement distribuées et que

$$\lim_{N\to\infty}\sqrt{N}E[X_1^{(N)}] = \delta \quad \text{et} \quad \lim_{N\to\infty} Var(X_1^{(N)}) = \sigma^2.$$

Dans ce nouveau cadre d'étude, on peut montrer que

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

où

$$H_N = \max_{0\leq i\leq j\leq N}\left(\sum_{k=i}^{j} X_k^{(N)}\right)$$

et

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

On montre que la fonction de répartition de la loi limite ξ_γ vérifie une équation intégrale, ce qui permet d'obtenir un développement asymptotique (cf théorème 4.9 page 55). Malheureusement ce développement n'est pas complètement explicite. Il est toutefois possible de donner un équivalent pour la queue de distribution (cf théorème 4.4 page 52) :

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

Ce travail a fait l'objet d'un article soumis à *Stochastic Processes and their Applications* .

Mots clé : *Score local, Mouvement brownien, Marches aléatoires.*

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

14th February 2003

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.

4.1 Introduction

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, when $\mathbb{E}(X_1)$ is "small".
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.
Biologists are often faced to the case where the expectation of X_n is "small". In practice they make use of Dembo and Karlin's approximation, which assumes that $\mathbb{E}(X_n) < 0$. However this approximation may be misleading. This motivates this study.
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 is closely connected to 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 :

$$\lim_{n\to+\infty} \mathbb{P}\left(H_n \leq \frac{\ln n}{\lambda} + x\right) = \exp\left(-K^* \exp\left(-\lambda x\right)\right) \tag{4.1.1}$$

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

In this paper 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 4.2). In this context, we prove that :

$$\frac{H_n}{\sqrt{n}} \xrightarrow[n\to\infty]{(d)} \sigma B_1^*, \tag{4.1.2}$$

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 explicitly defined as a series (cf Proposition 4.2).

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 :

$$\lim_{N\to+\infty} \sqrt{N}\mathbb{E}\left(X_1^{(N)}\right) = \delta \in \mathbb{R}, \quad \lim_{N\to+\infty} Var\left(X_1^{(N)}\right) = \sigma^2 > 0. \tag{4.1.3}$$

If the sequence $(X_k)_{k\geq 1}$ does not depend on N, then (4.1.3) is equivalent to : $\mathbb{E}(X_1) = 0$ and $Var(X_1) = \sigma^2$. This new setting includes thus the previous one.
Suppose from now that (4.1.3) holds.
Notice that $\mathbb{E}\left(X_1^{(N)}\right) \to 0$, when $N \to \infty$.
We prove (cf. proposition 4.5) that the analog of (4.1.2) is :

$$\frac{H_N^{(N)}}{\sqrt{N}} \underset{n\to\infty}{\overset{(d)}{\longrightarrow}} \sigma\xi_{\delta/\sigma}, \tag{4.1.4}$$

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

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 defined as the right hand side of (4.1.1).
- If $\mathbb{E}(X_1) > 0$, 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 $H_n \underset{n\to+\infty}{\sim} \mathbb{E}(X_1)\, n$.
- 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". 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 $(\sigma\xi_{\delta/\sigma})\sqrt{n}$.

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 4.9).
Let us introduce

$$\xi_\gamma(t) = \max_{0\leq u\leq t}\left\{B(u) + \gamma u - \min_{0\leq s\leq u}(B(s) + \gamma s)\right\}, \quad t \geq 0. \tag{4.1.5}$$

and

$$T_a = \inf\left\{t \geq 0; B(t) + \gamma t - \min_{0\leq s\leq t}(B(s) + \gamma s) > a\right\}, \quad a > 0. \tag{4.1.6}$$

Obviously $\xi_\gamma = \xi_\gamma(1)$.

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

$$\mathbb{E}\left[e^{-\lambda^2 T_a/2}\right] = \frac{\nu e^{\gamma a}}{\nu \cosh \nu a + \gamma \sinh \nu a}, \quad \lambda > 0. \tag{4.1.7}$$

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

$$\mathbb{P}\left(T_a < t\right) = \mathbb{P}\left(\xi_\gamma(t) > a\right), \quad \forall t \geq 0. \tag{4.1.8}$$

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

$$\mathbb{P}\left(\xi_\gamma(\alpha) > a\right) = \mathbb{P}\left(T_a < \alpha\right) = \mathbb{E}\left[e^{-T_a}\right]; \quad \forall a > 0. \tag{4.1.9}$$

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

$$1 - \mathbb{P}\left((\xi_\gamma(\alpha) \leq a\right) = \frac{\nu e^{\gamma a}}{\nu \cosh \nu a + \gamma \sinh \nu a}; \quad \forall a > 0. \tag{4.1.10}$$

As we said above, the distribution of ξ_γ is not easy to handle. So we investigate the tail of ξ_γ. We prove (cf Theorem 4.4) :

$$\mathbb{P}\left(\xi_\gamma \geq a\right) \underset{a \to \infty}{\sim} 2\sqrt{\frac{2}{\pi}} \frac{1}{a} e^{-(\gamma - a)^2/2}. \tag{4.1.11}$$

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 $t \to \infty$.
Now let us briefly describe the organization of the paper. In section 4.2, we study the convergence of H_n when n goes to infinity, the underlying random variables X_i being centered. In section 4.3, we investigate the asymptotic behaviour of H_n when the X_i have a bias depending on N. In this section we give a short introduction to our approach, state the results and detail only short proofs. The more technical proofs are postponed in section 4.4.

Acknowledgment. We would like to thank the referee for his interesting remarks and suggestions (in particular a direct proof of theorem 4.4).

4.2 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 :

$$S_0 = 0, \qquad S_k = \sum_{i=1}^{k} X_i; \qquad k \geq 1. \tag{4.2.1}$$

and the local score

$$H_n = \max_{0\le i\le j\le n}(S_j - S_i) = \max_{0\le i\le j\le n}(X_{i+1} + \ldots + X_j). \tag{4.2.2}$$

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

$$\begin{cases} t \mapsto H^{(N)}(t) \text{ is linear on each interval of the form } \left[\frac{j}{N}; \frac{j+1}{N}\right] \\ H^{(N)}\left(\frac{j}{N}\right) = \frac{1}{\sqrt{N}} H_j. \end{cases} \tag{4.2.3}$$

In this section the sequence $(X_n)_{n\ge 1}$ 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

$$\sigma^2 = \mathbb{E}_\nu(X_1^2) + 2\sum_{k=2}^{\infty} \mathbb{E}_\nu(X_1 X_k), \tag{4.2.4}$$

where ν is the invariant distribution of $(X_n)_{n\ge 0}$.
σ^2 is well defined for the series (4.2.4) is convergent ([Bil68, p. 166]).
We are now able to state the main result of this section :

Theorem 4.1 Let $(X_n)_{n\ge 1}$ be a sequence of random variables as above. Then the sequence of processes $\left(H^{(N)}(t),\ t \ge 0\right)$ converges in law to the process $(\sigma \max_{0\le u\le s} |B_u|, s \ge 0)$, as N tends to infinity.

Proof : We just outline the proof, the complete developments are given in section 4.4.1.
Let $B^{(N)}$ be the piecewise linear process defined by

$$B^{(N)}\left(\frac{k}{N}\right) = \frac{1}{\sigma\sqrt{N}} S_k; \qquad k \ge 0. \tag{4.2.5}$$

and

$$t \mapsto B^{(N)}(t) \text{ is linear on each interval of the form } \left[\frac{k}{N}, \frac{k+1}{N}\right] \tag{4.2.6}$$

It is well known ([Bil68]) that $\left(B^{(N)}(s), s \ge 0\right)$ converges to the standard Brownian motion. We easily check that $\left(H^{(N)}(s), s \ge 0\right)$ may be approached by a continuous function of $\left(B^{(N)}(s), s \ge 0\right)$ up to a remainder term R_N which converges to 0. This completes the proof of theorem 4.1.

□

An important application of theorem 4.1 is the convergence of the local score :

Proposition 4.2 1) $\frac{H_n}{\sqrt{n}}$ converges in distribution, as $n \to \infty$, to σB_1^*, where $B_1^* = \max_{0\leq u\leq 1}(|B_u|)$.
2) The cumulative distribution function (c.d.f) of B_1^* is :

$$\mathbb{P}(B_1^* \leq x) = \frac{2}{\pi}\sum_{k\in\mathbb{Z}} \frac{(-1)^k}{2k+1} \exp\left(-\frac{(2k+1)^2\pi^2}{8x^2}\right), \quad x \geq 0. \tag{4.2.7}$$

Proof : Theorem 4.1 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 (4.2.7) is classical and may be deduced from [BS96] (p.146) see 4.4.6 by a short calculus.

□

Remark 4.3 Theorem 4.1 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

$$\frac{T_a(H)}{a^2} \underset{a\to\infty}{\overset{(d)}{\longrightarrow}} \frac{1}{\sigma^2(B_1^*)^2}. \tag{4.2.8}$$

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

$$\mathbb{P}\left(\frac{T_{x\sqrt{N}}}{N} < t\right) \underset{N\to\infty}{\longrightarrow} \mathbb{P}\left(\sigma\sqrt{t}B_1^* > x\right) = \mathbb{P}\left(\frac{x^2}{\sigma^2(B_1^*)^2} < t\right). \tag{4.2.9}$$

Let $a = x\sqrt{N}$, (4.2.8) follows immediately.

□

4.3 Convergence in the non-centered case.

$(B_t; t \geq 0)$ 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 :

$$\lim_{N\to\infty} Var(X_1) = \sigma^2 > 0 \quad ; \quad \lim_{N\to\infty} \sqrt{N}\mathbb{E}(X_1) = \delta \in \mathbb{R}. \tag{4.3.1}$$

In this setting it is easy to prove (cf proposition 4.5) that $H_N/\sqrt{N}$ converges in distribution, when N goes to infinity, to $\sigma\xi_{\delta/\sigma}$, where

$$\xi_\gamma = \max_{0\leq u\leq 1}\left\{B(u) + \gamma u - \min_{0\leq s\leq u}(B(s) + \gamma s)\right\}. \tag{4.3.2}$$

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

$$\phi^{(\gamma)}(a) = e^{-\gamma a}\mathbb{P}(\xi_\gamma > a), \quad a \geq 0. \tag{4.3.3}$$

Let us briefly detail our approach. We state the main result (theorem 4.4) at the end of the subsection.
In section 4.2 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 4.6 and theorem 4.7) :

$$\phi^{(\gamma)}(a) = \frac{1}{a}\int_{[0,+\infty[^2} \mathbb{1}_{\{u\leq 1\}} \exp\left\{-\gamma t - \gamma^2 u/2\right\} \mu_a(u)\, F_t^{(\gamma)}(1-u, 1/a)\, du\, dt \tag{4.3.4}$$

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

$$F_t^{(\gamma)}(x,b) = \mathbb{E}\left(\mathbb{1}_{\{0\leq\tau_t\leq x,\ 0\leq B^*_{\tau_t}\leq 1/b\}} e^{-\gamma^2\tau_t/2}\right); \qquad x \geq 0,\ b \geq 0,\ t \geq 0. \tag{4.3.5}$$

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 (4.3.16) and (4.3.17)). 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 4.10),
- $F_t^{(\gamma)}$ is solution of an integral equation (cf (4.3.20)).

Moreover relation (4.3.20) yields to express $F_t^{(\gamma)}$ as sum of a series (Theorem 4.9). Unfortunately the coefficients are not explicit and are determined by a recursive algorithm.
However relation (4.3.20) is rich enough since we determine the decay rate of $a \to P(\xi_\gamma > a)$, $a \to \infty$. More precisely

Theorem 4.4 For all γ in $\mathbb{R}$:

$$\mathbb{P}(\xi_\gamma \geq a) \underset{a\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}} e^{-\gamma^2/2} \frac{1}{a} e^{\gamma a - a^2/2} = 2\sqrt{\frac{2}{\pi}} \frac{1}{a} e^{-(\gamma-a)^2/2}. \qquad (4.3.6)$$

Two proofs of Theorem 4.4 will be given. The first one is a consequence of Theorem 4.7 and is postponed in section 4.4.4. The second one suggested by the referee will be developed at the end of this section.

We now prove the main result mentioned in 3.1 (Theorems 4.7 and 4.9).
To help the reader we restrict ourself to short and easy proofs, the more technical points are postponed in the last section 4.4.

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 (4.3.1) 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

$$B^{(N)}\left(\frac{k}{N}\right) = \frac{1}{\sigma\sqrt{N}}(S_k - \mathbb{E}(S_k)) = \frac{1}{\sigma\sqrt{N}}(S_k - k\,a_N); \qquad k \geq 0. \qquad (4.3.7)$$

and

$$t \mapsto B^{(N)}(t) \text{ is linear on each interval of the form } \left[\frac{k}{N}, \frac{k+1}{N}\right] \qquad (4.3.8)$$

The process $(H^{(N)}(t), t \geq 0)$ is defined by the same procedure as in the centered case, i.e. formula (4.2.3). 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 4.5 1. Let $t > 0$. As N tends to ∞,

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

where

$$\xi_\gamma(t) = \max_{0 \le u \le t} \left\{ B(u) + \gamma u - \min_{0 \le s \le u} (B(s) + \gamma s) \right\}. \tag{4.3.9}$$

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 4.4.2 for a complete proof).

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

$$\xi_\gamma(t) \overset{(d)}{=} \sqrt{t} \xi_{\gamma\sqrt{t}}, \quad \text{for any } t > 0. \tag{4.3.10}$$

This leads us to determine the distribution of ξ_γ.

Proposition 4.6 For all $a > 0$ and $\gamma \in \mathbb{R}$, we set

$$\phi^{(\gamma)}(a) = e^{-\gamma a} \mathbb{P}(\xi_\gamma > a). \tag{4.3.11}$$

Then

$$\phi^{(\gamma)}(a) = \mathbb{E}\left[\mathbb{1}_{\{\tau_Z + T_a < 1\}} \exp\left\{ -\gamma Z - \frac{\gamma^2}{2}(\tau_Z + T_a) \right\} | B^*_{\tau_Z} < a \right], \quad \gamma \in \mathbb{R}, \tag{4.3.12}$$

where

- τ_t denotes the first time where the local time at 0 of Brownian motion $(B_t; t \ge 0)$ 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 $\frac{1}{a} e^{-x/a} \mathbb{1}_{\{x>0\}}$).
- $(B^*_u; u \ge 0)$ is the process : $B^*_u = \sup_{0 \le s \le u} |B_s|$, $u \ge 0$.
- for any $a > 0$, $(B_t; t \ge 0)$, 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 4.4.3.

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 :

$$\phi_\lambda^{(\gamma)}(a) = \mathbb{E}\left[\mathbb{1}_{\{\tau_Z + T_a < 1\}} \exp\left\{-\gamma Z - \frac{\lambda^2}{2}(\tau_Z + T_a)\right\} | B^*_{\tau_Z} < a\right], \quad \lambda \in \mathbb{R}, \tag{4.3.13}$$

In our approach it is not more difficult to deal with $\phi_\lambda^{(\gamma)}$ instead of $\phi^{(\gamma)}$. Formula (4.3.12) 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 (4.3.12) is the following :

Theorem 4.7 Let $\lambda \in \mathbb{R}$ be fixed, then for any $a > 0$

$$\phi_\lambda^{(\gamma)}(a) = \frac{1}{a} \int_{[0,+\infty[^2} \mathbb{1}_{\{u \le 1\}} \exp\left\{-\gamma t - \lambda^2 u/2\right\} \mu_a(u)\, F_t^{(\lambda)}(1-u, 1/a)\, du\, dt \tag{4.3.14}$$

where

$$F_t^{(\lambda)}(x,b) = \mathbb{E}\left(\mathbb{1}_{\left\{0 \le \tau_t \le x,\ 0 \le B^*_{\tau_t} \le 1/b\right\}} e^{-\lambda^2 \tau_t/2}\right); \qquad x \ge 0,\ b \ge 0,\ t \ge 0, \tag{4.3.15}$$

and μ_a is the density function of T_a :

$$\mu_a(t) = \frac{1}{a^2} \mu_1\left(\frac{t}{a^2}\right), \tag{4.3.16}$$

and

$$\mu_1(t) = \frac{1}{\sqrt{2\pi}\, t^{3/2}} \sum_{k \in \mathbb{Z}} \left(-1 + \frac{(1+2k)^2}{t}\right) \exp -\frac{(1+2k)^2}{2t}. \tag{4.3.17}$$

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

$$\mu_1(t) = \frac{d}{dt} \sum_{n=-\infty}^{\infty} (-1)^n e^{-(n^2\pi^2 t)/2}. \tag{4.3.18}$$

Remark 4.8 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 4.7 : The random variables involved in equation (4.3.12) being independent, we have :

$$\phi_\lambda^{(\gamma)}(a) = \frac{1}{a}\int_{[0,+\infty[^2} \mathbb{E}\left[\mathbb{1}_{\{\tau_t+u<1\}} e^{\left(-\gamma t - \frac{\lambda^2}{2}(\tau_t+u)\right)} \mid B^*_{\tau_t} < a\right] \mu_a(u) e^{-t/a} \mathrm{d}u\, \mathrm{d}t, \tag{4.3.19}$$

where μ_a denotes the density function of T_a.
Using Remark 4.8, equation (4.3.14) 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 4.9 Let $\lambda \in \mathbb{R}$ and $t \ge 0$ be two fixed parameters.

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

$$F_t^{(\lambda)}(x,a) = F_t^{(\lambda)}(x,0) - t\left(A^{(\lambda)} F_t^{(\lambda)}\right)(x,a), \quad (x,a) \in \mathbb{R}_+^2, \tag{4.3.20}$$

with

$$\left(A^{(\lambda)}\psi\right)(x,a) = \int_{[0,+\infty[^2} \mathbb{1}_{\{u\le a, y\le x\}} \mu_{1/u}^{(2)}(y)\, e^{-\lambda^2 y/2} \psi(x-y,u)\, \mathrm{d}y\, \mathrm{d}u, \tag{4.3.21}$$

$$F_t^{(\lambda)}(x,0) = \frac{t}{\sqrt{2\pi}} \int_0^x \exp\left(-\frac{\lambda^2 z}{2} - \frac{t^2}{2z}\right) \frac{\mathrm{d}z}{z^{3/2}}, \tag{4.3.22}$$

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

Recall (cf [BPY01]) that :

$$\mu_1^{(2)}(t) = \frac{d}{dt}\left(\frac{8\sqrt{2}}{\sqrt{\pi}t^{3/2}} \sum_{n=1}^{+\infty} n^2 e^{-2n^2/t}\right). \tag{4.3.23}$$

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

$$F_t^{(\lambda)}(x,a) = \sum_{k=0}^{+\infty} (-1)^k\, t^k\, \alpha_t^{(k)}(x,a), \tag{4.3.24}$$

where

$$\alpha_t^{(0)}(x,a) = F_t^{(\lambda)}(x,0), \tag{4.3.25}$$

$$\alpha_t^{(k+1)}(x,a) = \left(A^{(\lambda)}\alpha_t^{(k)}\right)(x,a). \tag{4.3.26}$$

The convergence of (4.3.24) 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

$$\mathbb{P}(\xi_\gamma > a) = \frac{e^{\gamma a}}{a} \sum_{k\geq 0} (-1)^k \int_{[0,+\infty[^2} \mathbb{1}_{\{u\leq 1\}} e^{-\gamma t - \lambda^2 u/2} \mu_a(u) t^k \alpha_t^{(k)}(1-u, 1/a) \mathrm{d}u\, \mathrm{d}t. \tag{4.3.27}$$

Proof : see section 4.4.5.
Making use of Theorem 4.9 (especially formula (4.3.20)) 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 (4.3.28). As we notice in 3.1, θ_t is unknown, therefore (4.3.28) is interesting.
As formula (4.3.14) 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 (4.3.24) of $F_t^{(\lambda)}$ cannot be deduced from it. This justifies our choice : $F_t^{(\lambda)}$ is the right parameter.

Proposition 4.10 Let $t > 0$. The random variable $(\tau_t, B^*_{\tau_t})$ has a density function θ_t. Moreover θ_t verifies :

$$\theta_t(x, a) = \frac{t}{a^2} \int_{[0,+\infty[^2} \mathbb{1}_{[0,x]\times[0,a]}(y, b) \mu_a^{(2)}(x - y) \theta_t(y, b)\, \mathrm{d}y\, \mathrm{d}b. \tag{4.3.28}$$

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 (4.3.20), we get :

$$f([0, x] \times [a, +\infty[) = t \int_0^{1/a} \mathrm{d}u \left(\int_0^x \mu_{1/u}^{(2)}(y) f([0, x - y] \times [0, 1/u]) \mathrm{d}y \right). \tag{4.3.29}$$

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

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

The new relation obtained by setting $v = 1/a$ in (4.3.29) 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 4.4 suggested by the referee.

Proof of Theorem 4.4 : 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

$$\mathbb{P}(\xi_\gamma > a) = \mathbb{P}(T_a < 1). \tag{4.3.30}$$

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

$$\mathbb{E}\left[e^{-\frac{\lambda^2}{2}T_a}\right] = \frac{\nu e^{\gamma a}}{\nu\cosh\nu a + \gamma\sinh\nu a}. \tag{4.3.31}$$

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 (4.3.30), we explicit the asymptotic behaviour of $\mathbb{P}(\xi_\gamma > a)$, $a \to \infty$.
1) By an easy computation we have :

$$\begin{aligned}\mathbb{E}\left[e^{-\frac{\lambda^2}{2}T_a}\right] &= \frac{2\nu e^{(\gamma-\nu)a}}{\gamma+\nu}\frac{1}{1+\frac{\nu-\gamma}{\gamma+\nu}e^{-2\nu a}},\\ &= 2\nu e^{\gamma a}\left(\sum_{k\ge 0}(-1)^k\frac{(\nu-\gamma)^k}{(\gamma+\nu)^{k+1}}e^{-(2k+1)\nu a}\right).\end{aligned}$$

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

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

This yields to

$$\mathbb{E}\left[e^{-\frac{\lambda^2}{2}T_a}\right] = 2e^{\gamma a}\left(\sum_{k\ge 0}(-1)^k\int_0^{+\infty}\nu e^{-\nu(t+(2k+1)a)-\gamma t}L_k(2\gamma t)\mathrm{d}t\right). \tag{4.3.32}$$

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 (4.3.32) : T_a has a density ϕ_a and

$$\phi_a(x) = \sqrt{\frac{2}{\pi}}e^{\gamma a}\frac{e^{-\gamma^2 x/2}}{x^{5/2}}\left(\sum_{k\geq 0}\psi_{a,k}(x)\right), \tag{4.3.33}$$

where

$$\psi_{a,k}(x) = (-1)^k\int_0^{+\infty}\left((t+(2k+1)a)^2 - x\right)e^{-\gamma t}L_k(2\gamma t)e^{-(t+(2k+1)a)^2/2x}\mathrm{d}t. \tag{4.3.34}$$

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 :

$$\psi_{a,0}(x) \underset{a\to\infty}{\overset{u}{\sim}} xae^{-a^2/2x}. \tag{4.3.35}$$

We set $t+a = \sqrt{x}u$ in (4.3.34) (with $k=0$) :

$$\psi_{a,0}(x) = x^{3/2}e^{\gamma a}\int_{a/\sqrt{x}}^{+\infty}(u^2-1)e^{-u^2/2}e^{-\gamma\sqrt{x}u}\mathrm{d}u. \tag{4.3.36}$$

But

$$\left(-ue^{-u^2/2}\right)' = (u^2-1)e^{-u^2/2}, \tag{4.3.37}$$

then integrating by part in (4.3.36) 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 $x\in[0;1]$, then

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

(4.3.35) follows immediately.

3) We claim that

$$\sum_{k\geq 0} \psi_{a,k}(x) \underset{a\to\infty}{\overset{u}{\sim}} \psi_{a,0}(x). \tag{4.3.38}$$

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

$$(t+(2k+1)a)^2 \geq a^2 \geq 1 > x; \quad \forall t \geq 0. \tag{4.3.39}$$

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

$$\mid L_k(x) \mid \leq e^{x/2}; \quad x \geq 0. \tag{4.3.40}$$

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 (4.3.37) 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 $a > 1$,

$$\mid \psi_{a,k}(x) \mid \leq \left(x\,a\,e^{-a^2/2x}\right)\left((2k+1)e^{-2k^2}e^{-2a^2}\right). \tag{4.3.41}$$

This demonstrates (4.3.38).

4) Let us end the proof of Theorem 4.4. Using both (4.3.30), (4.3.33), (4.3.35) and (4.3.38) 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 4.4.

□

4.4 Technical proofs

This section is devoted to the proofs of Theorems 4.1, 4.9, 4.4, Propositions 4.5, 4.6 and formulae (4.2.7), (4.3.17), (4.3.18) and (4.3.23).

4.4.1 Proof of Theorem 4.1

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 (4.2.4).
Given an integer $N \geq 0$, we consider the piecewise linear process $B^{(N)}(t)$

$$\begin{cases} B^{(N)}\left(\frac{k}{N}\right) = \frac{1}{\sigma\sqrt{N}}\left(S_k - \mathbb{E}(S_k)\right) = \frac{1}{\sigma\sqrt{N}}\, S_k; \qquad k \geq 0, \\ t \mapsto B^{(N)}(t) \text{ is linear on each interval of the form } \left[\frac{k}{N}, \frac{k+1}{N}\right]. \end{cases} \tag{4.4.1}$$

Our approach is based on the two following results.

Theorem 4.11 (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 4.12 (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 4.1: 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 :

$$H^{(N)}(t) = \frac{1}{\sqrt{N}}\left\{H_{[Nt]} + (Nt - [Nt])\left(H_{[Nt]+1} - H_{[Nt]}\right)\right\}, \quad t \geq 0. \tag{4.4.2}$$

1. Relation (4.4.1) implies :

$$S_k = \sigma\sqrt{N}\, B^{(N)}\left(\frac{k}{N}\right). \tag{4.4.3}$$

Then

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

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

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

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

$$H^{(N)}(t) = \sigma\left(\max_{0\le u\le v\le \frac{[Nt]}{N}} \left\{ B^{(N)}(v) - B^{(N)}(u)\right\} + R_N(t)\right), \tag{4.4.4}$$

$$\begin{aligned} R_N(t) = (Nt - [Nt]) \Bigg(& \max_{0\le u\le v\le \frac{[Nt]+1}{N}} \left\{ B^{(N)}(v) - B^{(N)}(u)\right\} \\ & - \max_{0\le u\le v\le \frac{[Nt]}{N}} \left\{ B^{(N)}(v) - B^{(N)}(u)\right\}\Bigg). \end{aligned} \tag{4.4.5}$$

2. We apply theorems 4.11 and 4.12 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 (4.4.5) (resp. (4.4.4)) 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 $t \in [0,T]$:

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

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

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

$$\left(\underline{H}^{(N)}(t), 0\le t\le T\right) \xrightarrow[N\to\infty]{a.s.} (\sigma \max_{0\le u\le v\le t} \{\underline{B}(v)-\underline{B}(u)\}\,;\, 0\le t\le T). \tag{4.4.6}$$

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 4.1. □

4.4.2 Proof of Proposition 4.5

This proof is similar to the previous one (see section 4.4.1) Let $H^{(N)}$ be the piecewise linear function defined by (4.4.2). The equation (4.4.3) has to be replaced by

$$\frac{1}{\sqrt{N}}S_k = \sigma_N B^{(N)}\left(\frac{k}{N}\right) + \frac{k a_N}{\sqrt{N}} = \sigma_N B^{(N)}\left(\frac{k}{N}\right) + \frac{k}{N}(\sqrt{N}a_N), \tag{4.4.7}$$

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

$$\frac{H_{[N\,t]}}{\sqrt{N}} = \max_{0\le u\le v\le \frac{[N\,t]}{N}} \left\{\sigma_N B^{(N)}(v) + v\,(\sqrt{N}a_N) - \sigma_N B^{(N)}(u) - u\,(\sqrt{N}a_N)\right\} \tag{4.4.8}$$

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

4.4.3 Proof of Proposition 4.6

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 $\phi^{(\gamma)}$:

$$\phi^{(\gamma)}(a) = \mathbb{E}\left[\mathbb{1}_{\{T_a^*<1\}} \exp\left(-\gamma\, L^0_{T_a^*} - \frac{\gamma^2}{2}\,T_a^*\right)\right], \tag{4.4.9}$$

Let f be a Borel bounded function, we have :

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

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

$$\mathbb{E}\left[f\left(\xi^{(\gamma)}\right)\right] = \mathbb{E}\left[f\left(\max_{0\le u\le 1}\left(B_u - \min_{0\le s\le u} B_s\right)\right)\exp\left\{\gamma B_1 - \frac{\gamma^2}{2}\right\}\right]. \tag{4.4.10}$$

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

$$\mathbb{E}\left[f\left(\xi^{(\gamma)}\right)\right] = \mathbb{E}\left[f\left(\max_{0\le s\le 1}|B_s|\right)\exp\left\{\gamma\left(|B_1| - L_1^0\right) - \frac{\gamma^2}{2}\right\}\right]. \tag{4.4.11}$$

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 (4.4.11) reduces to :

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

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 :

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

This shows (4.4.9).

2) We are now able to prove (4.3.12).
The proof is based on decomposition of Brownian path ([Val91b], prop 4). Let us recall this decomposition :
For $a > 0$. Define

$$g = \sup\left\{t\le T_a^*,\ B_t = 0\right\}.$$

Then

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

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

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

We decompose T_a^* as the sum of g and $T_a^* - \gamma$, (4.3.12) is a straightforward consequence of (4.4.9). □

4.4.4 Second proof of Theorem 4.4

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

Lemma 4.13 Let ψ be the function :

$$\psi(v) = \int_{\mathbb{R}_+} e^{-\gamma y} F_y\left(\frac{v}{1+v}, 0\right) \mathrm{d}y; \quad v > 0. \tag{4.4.12}$$

Then

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

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

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

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 ,

$$\psi(v) = \frac{1}{\sqrt{2\pi}} \int_0^{v/1+v} \frac{1}{z^{1/2}} e^{-\lambda^2 z/2} \int_0^{+\infty} u e^{-(u^2/2+\gamma u\sqrt{z})} \mathrm{d}u\, \mathrm{d}z. \tag{4.4.13}$$

We have :

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

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

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

$$\psi_2(v) = \frac{1}{\sqrt{2\pi}} \int_0^{v/1+v} \frac{1}{z^{1/2}} e^{-\lambda^2 z/2} \left(\int_0^{+\infty} u e^{-u^2/2} \mathrm{d}u\right) \mathrm{d}z, \tag{4.4.14}$$

$$\psi_3(v) = \frac{1}{\sqrt{2\pi}} \int_0^{v/1+v} \frac{1}{z^{1/2}} e^{-\lambda^2 z/2} \left(\int_0^{+\infty} u e^{-u^2/2} \rho(\gamma u \sqrt{z}) du \right) dz. \tag{4.4.15}$$

Clearly

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

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

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

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 4.4 : Let us recall the expression of $\phi_\lambda^{(\gamma)}$ given in equation (4.3.14).

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

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

$$\rho_1(a) = \frac{1}{a} \int_{\mathbb{R}_+^2} \mathbb{1}_{\{u \leq 1\}} \exp\left\{-\gamma y - \frac{\lambda^2}{2} u\right\} \mu_a(u) F_y^{(\lambda)}(1-u, 0) du \, dy. \tag{4.4.16}$$

   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 (4.4.16).

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

$$\rho_2(a) = \frac{2a^2}{\sqrt{2\pi}} \int_{\mathbb{R}_+^2} \mathbb{1}_{\{u \leq 1\}} e^{-\gamma y - \lambda^2 u/2 - a^2/2u} F_y^{(\lambda)}(1-u, 0) \frac{du}{u^{5/2}} \, dy. \tag{4.4.17}$$

We use the explicit form of μ_a given by equation (4.3.17), the scaling property, and (4.3.16) 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

$$R(u,a)=\sum_{k\in\mathbb{Z}}\left(-1+a^2\frac{(1+2k)^2}{u}\right)\exp\left\{-a^2\frac{(1+2k)^2}{2u}\right\}. \quad (4.4.18)$$

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 $a\geq 1$ and $u\leq 1$, we have :

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

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

This yields

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

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 :

$$\rho_2(a) = \frac{2a^2}{\sqrt{2\pi}} e^{-a^2/2} \int_0^{+\infty} e^{-a^2 v/2} e^{-\lambda^2/2(1+v)} \sqrt{1+v}\, \psi(v)\mathrm{d}v, \quad (4.4.19)$$

Let us set $u = a^2 v/2$ in equation (4.4.19), 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 4.13 implies that :

$$\rho_2(a) = \frac{4}{\sqrt{2\pi}} e^{-a^2/2} \left[\frac{2\sqrt{2}}{\sqrt{2\pi}} \frac{1}{a} \int_0^{+\infty} e^{-u} e^{-\lambda^2 a^2/2(a^2+2u)} \sqrt{u}\mathrm{d}u + \rho_3(a)\right], \quad (4.4.20)$$

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 (4.4.20) 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 (4.3.6).

□

4.4.5 Proof of Theorem 4.9

We divide the proof into two steps.
1) Let $F_t^{(\lambda)}$ be the function defined by (4.3.15) :

$$F_t^{(\lambda)}(x,b) = \mathbb{E}\left(\mathbb{1}_{\{0\leq\tau_t\leq x,\ 0\leq B^*_{\tau_t}\leq 1/b\}} e^{-\lambda^2\tau_t/2}\right), \quad x\geq 0, b\geq 0, \qquad (4.4.21)$$

Here λ and t are fixed. We have :

$$F_t^{(\lambda)}(x,b) = \mathbb{E}\left(\mathbb{1}_{\{0\leq\tau_t\leq x\}} e^{-\lambda^2\tau_t/2}\right) - \mathbb{E}\left(\mathbb{1}_{\{0\leq\tau_t\leq x,\ B^*_{\tau_t}>1/b\}} e^{-\lambda^2\tau_t/2}\right). \qquad (4.4.22)$$

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 4.8), we get :,

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

We set $v = 1/u$, (4.3.20) follows immediately since we have already established (4.3.15) in the proof of Lemma 4.13.
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

$$0 \leq F_t^{(\lambda)}(x,0) \leq 1. \tag{4.4.23}$$

Let us consider the series

$$\Lambda_t(x,a) = \sum_{k=0}^{+\infty} (-1)^k\, t^k \alpha_t^{(k)}(x,a), \tag{4.4.24}$$

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

$$\max_{x\geq 0, y\leq a} |\alpha_t^{(k)}(x,y)| \leq \frac{a^k}{k!} \max_{x\geq 0, y\leq a} |\alpha_t^{(0)}(x,y)| \leq \frac{a^k}{k!}. \tag{4.4.25}$$

We check (4.4.25) by induction on n.
If $n = 0$, obviously (4.4.25) holds. We suppose that (4.4.25) is verified for n and we prove that (4.4.25) 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

$$\max_{x\geq 0, y\leq a} |A^{(\lambda)}\psi(x,y)| \leq \int_0^a \left(\max_{x\geq 0,\, u_1\leq u} |\psi(x,u_1)| \right) \mathrm{d}u. \tag{4.4.26}$$

Consequently (4.4.25) 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 (4.4.24) 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 (4.4.26) 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 (4.3.27). □

4.4.6 Proof of formula (4.2.7).

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) :

$$\mathbb{P}\left(B_1^* < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} \sum_{k \in \mathbb{Z}} \left(e^{-\frac{(y+4kx)^2}{2}} - e^{-\frac{(y+2x+4kx)^2}{2}} \right) dy. \qquad (4.4.27)$$

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

$$\frac{1}{\sqrt{\pi t}} \sum_{k \in \mathbb{Z}} e^{-\frac{(v+k)^2}{t}} = \sum_{k \in \mathbb{Z}} \cos\left(2k\pi v\right) e^{-k^2\pi^2 t}, \quad v \in \mathbb{R},\ t > 0. \qquad (4.4.28)$$

Setting $v = y/4x$ and $t = 1/8x^2$,(4.4.28) becomes :

$$\frac{4x}{\sqrt{2\pi}} \sum_{k \in \mathbb{Z}} e^{-\frac{(4kx+y)^2}{2}} = \sum_{k \in \mathbb{Z}} \cos\left(k\pi y/2x\right) e^{-\frac{k^2\pi^2}{8x^2}}. \qquad (4.4.29)$$

Then

$$\frac{1}{\sqrt{2\pi}} \sum_{k \in \mathbb{Z}} e^{-\frac{(4kx+y)^2}{2}} = \frac{1}{4x} \sum_{k \in \mathbb{Z}} \cos\left(k\pi y/2x\right) e^{-\frac{k^2\pi^2}{8x^2}}. \qquad (4.4.30)$$

Similarly, setting $v = (y + 2x)/4x$ and $t = 1/8x^2$ in (4.4.28), we obtain :

$$\frac{1}{\sqrt{2\pi}} \sum_{k \in \mathbb{Z}} e^{-\frac{(4kx+2x+y)^2}{2}} = \frac{1}{4x} \sum_{k \in \mathbb{Z}} (-1)^k \cos\left(k\pi y/2x\right) e^{-\frac{k^2\pi^2}{8x^2}}. \qquad (4.4.31)$$

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

$$\mathbb{P}(B_1^* < x) = \frac{2}{\pi} \sum_{k \in \mathbb{Z}} \frac{(-1)^k}{2k+1} e^{-\frac{(2k+1)^2\pi^2}{8x^2}}. \qquad (4.4.32)$$

4.4.7 Proof of formula (4.3.17).

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

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

$$\mu_1(t) = \frac{1}{\sqrt{2\pi} t^{3/2}} \sum_{k \in \mathbb{Z}} \left(-1 + \frac{(1+2k)^2}{t}\right) \exp - \frac{(1+2k)^2}{2t}. \qquad (4.4.33)$$

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

$$P\left(T_a^{(x)} \in dt\right) = \frac{a}{x} \Psi_x^{(a)}(t) \mathbb{1}_{\{t \geq 0\}}\, dt = \varphi_x^{(a)}(t) \mathbb{1}_{\{t \geq 0\}}\, dt \qquad (4.4.34)$$

where

$$\Psi_x^{(a)}(t) = \frac{1}{\sqrt{2\pi t^3}} \sum_{k\in\mathbb{Z}} (a - x + 2ka) \exp -\frac{(a-x+2ka)^2}{2t}. \tag{4.4.35}$$

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

$$\mu_a(t) = \lim_{x\to 0} \varphi_x^{(a)}(t) = \lim_{x\to 0} \frac{a}{x} \left(\Psi_x^{(a)}(t) - \Psi_0^{(a)}(t) \right) \tag{4.4.36}$$

Differentiating term by term, we obtain :

$$\mu_a(t) = \frac{a}{\sqrt{2\pi}\, t^{3/2}} \sum_{k\in\mathbb{Z}} \left(-1 + \frac{a^2(1+2k)^2}{t} \right) \exp -\frac{a^2(1+2k)^2}{2t} \tag{4.4.37}$$

□

4.4.8 Proof of (4.3.18).

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

$$\sum_{k\in\mathbb{Z}} \varphi(a+2kb) = \frac{\pi}{b} \sum_{k\in\mathbb{Z}} f\left(\frac{k\pi}{b}\right) \exp\left(\frac{ik\pi a}{b}\right). \tag{4.4.38}$$

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 (4.4.38), 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 $t = \dfrac{4}{u\pi^2}$:

$$\sum_{k\in\mathbb{Z}}(-1)^k e^{-k^2\pi^2 u/2} = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{u}}\sum_{k\in\mathbb{Z}}\exp -\frac{(2k-1)^2}{2u}. \qquad (4.4.39)$$

Differentiating in respect to u, we obtain (4.3.18). □

4.4.9 Proof of formula (4.3.23).

We keep the notations introduced in the beginning of 4.4.7.
Let us recall that $\mu_1^{(2)}$ is the density function of $Z = T_1^{(0)} + \tilde{T}_1^{(0)}$, $\tilde{T}_1^{(0)}$ 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 (4.3.23). □

Bibliography

[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] P. Biane, J. Pitman, and M. 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] H. 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 edition (1944).

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

[Wil76] D. 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 5

Comparaisons de trois approximations pour le score local lorsque $E(X) \simeq 0$

Contents

Résumé :

Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires. Le score local de (X) est défini par $H_n = \max_{0\leq i\leq j\leq n}(X_i + \ldots + X_j)$.

Il existe différentes approximations pour la fonction de répartition de H_n : l'approximation par les valeurs extrêmes de Dembo et Karlin et deux autres présentées dans le chapitre précédent qui sont fondées sur une approche par le mouvement brownien. Elles ont des domaines de validité différents.

En effet, l'approximation de Karlin (équation (4.1.1), page 46) est valable lorsque $E(X) < 0$, l'approximation brownienne ((4.1.2) page 46) suppose que

l'espérance est nulle, tandis que la dernière approximation (donnée par la formule (4.1.11), page 48) n'est utilisable que pour la queue de distribution.

Lorsque la moyenne est négative mais très proche de 0, il est utile de se demander quelle est l'approximation qui donnera les meilleurs résultats.

Ce chapitre a donc pour but de comparer les différents moyens d'approcher la distribution du score local lorsque la moyenne de la suite considérée est proche de 0 et surtout d'essayer de déterminer un domaine de validité pour chacune d'elles.

Keywords : score local, mouvement brownien, comportement asymptotique.

AMS 1991 Subject classifications

60G50

5.1 Introduction

On considère $(X_i)_{i\geq 1}$ une suite de variables aléatoires qui modélise le score de chaque élément de la séquence considérée. On peut alors définir le score global de cette suite par :

$$S_k = X_1 + \ldots + X_k, \quad S_0 = 0. \tag{5.1.1}$$

Le score local s'exprime à l'aide de la suite $(S_k)_{k\geq 0}$ de la manière suivante :

$$H_n = \max_{0\leq i\leq j\leq n} (S_j - S_i) = \max_{0\leq j\leq n} \left(S_j - \min_{0\leq i\leq j} S_i \right) \tag{5.1.2}$$

Dans le chapitre précédent, nous avons vu deux approximations de la distribution de H_n. Il existe également une approximation établie par Karlin et Dembo [DK92] qui repose sur une approche par les valeurs extrêmes. Les hypothèses d'application de ces trois approximations sont théoriquement différentes, mais il est intéressant d'oberver pratiquement leurs différences et c'est l'objet de ce chapitre.

Après avoir rappelé les différentes approximations (cf. paragraphe 5.2) ainsi que leur domaine de validité théorique, nous présenterons les résultats des tests que nous avons effectués dans le but de mettre en évidence le domaine de validité numérique de chacune d'entre elles. Les programmes nécessaires aux différents calculs ont été faits à l'aide du logiciel matlab.

5.2 Les différentes approximations

Nous avons vu dans le chapitre 3 qu'il était utile de connaître la distribution du score local pour estimer la significativité statistique des résultats obtenus en utilisant cet outil. Mais le score local est un objet complexe et on a difficilement accès à sa distribution. Daudin et Mercier [DM99] ont obtenu une formule exacte pour $\mathbb{P}(H_n < x)$ en utilisant une approche par les chaînes de Markov. Ainsi $\mathbb{P}(H_n < x)$ s'exprime à l'aide d'un vecteur P_n de longueur x donné par :

$$P_n = P_0 \Pi^n$$

où $P_0 = (1, 0, \ldots, 0)$ est de longueur x et π est la matrice de transition d'une chaîne de Markov à x états. La distribution du score local est alors donnée par :

$$\mathbb{P}(H_n \geq x) = P_n(x). \tag{5.2.1}$$

En pratique cette formule n'est utilisable que pour x et n pas trop grands, sinon les temps de calcul à mettre en oeuvre pour accéder à la distribution deviennent trop importants. Ces limitations sur ces deux données posent quelques problèmes. En effet, on s'intéresse souvent à de longues séquences

(n grand) et aux événements rares, donc la queue de distribution (c'est-à-dire x élevé). Il est donc important de disposer d'approximations pour n et x grands.

Dans le cas d'une espérance négative pour la suite (X), Dembo et Karlin ont montré que ([DK92]) :

$$\lim_{n\to+\infty} \mathbb{P}\left(H_n \leq \frac{\ln n}{\lambda} + x\right) = \exp\left(-K^* \exp\left(-\lambda x\right)\right) \tag{5.2.2}$$

où K^* et λ dépendent uniquement de la loi de probabilité de X_1.

En fait, ce résultat est vrai uniquement pour des variables continues, dans le cas discret les mêmes auteurs obtiennent deux bornes. Plus récemment Bacro, Daudin, Mercier et Robin [BDMR02] ont donné de meilleures bornes qui sont :

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

où R, δ et θ ne dépendent que de la distribution de X_1.

Nous avons vu dans le chapitre précédent, dans le cas où la suite considérée (X) est centrée, que la loi limite du score local est une fonctionnelle brownienne :

$$\mathbb{P}\left(\frac{H_n}{\sqrt{n}} \geq a\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(\sigma B_1^* \geq a\right), \tag{5.2.3}$$

où $B_1^* = \max_{0\leq u\leq 1} |B_u|$ et $(B_u, u \geq 0)$ désigne un mouvement brownien standard issu de 0.

Enfin, on dispose d'une troisième approximation valable pour la queue de distribution uniquement, mais quelle que soit la moyenne de la suite considérée.

$$\mathbb{P}\left(H_n \geq a\right) \underset{a\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}}\frac{\sigma\sqrt{n}}{a} \exp -\frac{(\delta_n - a/\sqrt{n})^2}{2\sigma^2}, \tag{5.2.4}$$

où $\delta_n = \sqrt{n}\mathbb{E}(X_1)$.

Finalement, on a donc accès à trois approximations et à une expression exacte. Nous proposons donc plusieurs comparaisons des ces approximations en utilisant la valeur exacte comme valeur de référence. Ces approximations ont été conduites pour différentes valeurs de la moyenne $\mathbb{E}\left(X_1\right)$ de manière à mettre en évidence les différences de comportements de ces approximations. Le but de ce travail est bien évidemment de donner des indications pour le choix de la bonne approximation.

	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	6.7931e-07
	0.5407	0.0715	0.0039	8.2630e-05	5.5331e-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

TAB. 5.1 – Encadrement de la valeur exacte du score local par l'approximation brownienne

Lorsque la moyenne est nulle, la meilleure approximation est (5.2.3). De même lorsque la moyenne est nettement négative, il est bon d'utiliser l'approximation de Karlin et Dembo (5.2.2). Nous nous sommes donc concentrés sur l'étude de la transition de phase entre ces deux limites.

5.3 Quelques remarques préliminaires

5.3.1 Sur l'approximation brownienne

En effectuant les tests numériques dont un résumé est présenté dans le paragraphe 5.4, il semble qu'un encadrement pour la fonction de répartition se dégage. En effet sur tous les exemples traités, on peut observer l'inégalité suivante :

$$\mathbb{P}\left(B_1^* \geq (x+K)/\sigma\sqrt{n}\right) \leq \mathbb{P}\left(H_n \geq x\right) \leq \mathbb{P}\left(B_1^* \geq x/\sigma\sqrt{n}\right) \qquad (5.3.1)$$

où K désigne l'accroissement maximal, c'est-à-dire $K = \max_{\omega\in\Omega} X(\omega)$.

Ce résultat reste une conjecture mais nous n'avons observé aucun contre-exemple numérique.

Le tableau 5.1 présente cette inégalité. Il se lit de la manière suivante : à l'intersection de la ligne 1 (resp. 2 ou 3) et de la colonne i, on va lire la probabilité que le score local d'une suite de longueur 100 (resp. 900 ou 10000) dépasse le seuil $i*\sqrt{100}$ (resp. $i*\sqrt{900}$ ou $i*\sqrt{10000}$). On lit alors au centre la valeur exacte obtenue à l'aide de la formule (5.2.1), au dessus l'approximation brownienne et en-dessous l'approximation brownienne corrigée.

Ces résultats numériques permettent de visualiser la convergence du score local normalisé $H_n/\sqrt{n}$ vers B_1^*. En effet, la valeur centrale se rapproche

visiblement de la valeur supérieure lorsque l'on s'intéresse à des séquences plus longues.

Lorsque les séquences sont longues mais que les incréments ne sont pas de moyenne nulle, l'approximation brownienne n'est plus justifiée théoriquement. Dans le chapitre précédent, nous avons vu que l'on pouvait utiliser une autre approximation mais qui ne concerne que la queue de distribution.

5.3.2 Sur l'approximation de la queue de distribution

Il est difficile de donner un seuil théorique à partir duquel cette approximation peut être validée. En fait, même si cette approximation est valable en théorie quelle que soit la valeur de l'espérance des variables considérées, le seuil de validation, lui, en dépend fortement.

Rappelons comment on obtient cette approximation. On pose tout d'abord $\delta = \sqrt{n}\mathbb{E}(X_1)$, on montre alors que

$$\mathbb{P}\left(\frac{H_n}{\sqrt{n}} \geq x\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(\sigma\xi_{\delta/\sigma} \geq x\right).$$

Dans un second temps, on montre que l'on a un équivalent de la queue de distribution de la variable limite $\sigma\xi_{\delta/\sigma}$ de la forme :

$$\mathbb{P}(\sigma\xi_{\delta/\sigma} \geq x) \underset{x\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}}\frac{x}{\sigma}\exp -\frac{\delta - a}{2\sigma^2}.$$

Il y a donc deux approximations dans ce résultat : la première consiste à approcher le score local par sa limite et la seconde à identifier la queue de distribution à son équivalent.

La figure 5.1 page 83 propose un exemple où l'approximation par la queue de distribution est bonne très vite. Les calculs ont été effectués en utilisant une loi de Bernoulli $\mathcal{B}(0.505)$ pour les incréments et en travaillant sur une séquence de longueur 900. Nous présentons sur cette figure deux fonctions :

$$x \mapsto \mathbb{P}\left(H_{900} > x\right) \quad \text{et} \quad x \mapsto \sqrt{\frac{2}{\pi}}\frac{60}{x}\ \exp\frac{-(0.3 - x/30)^2}{2}.$$

La première formule est déterminée avec la formule exacte (5.2.1) et la seconde utilise l'approximation par la queue de distribution donnée dans l'équation (5.2.4).

Un zoom sur la figure 5.1 donne la figure 5.2. Le seuil approximatif de validité serait $900 * 0.01 + 5 * 30 = 159$

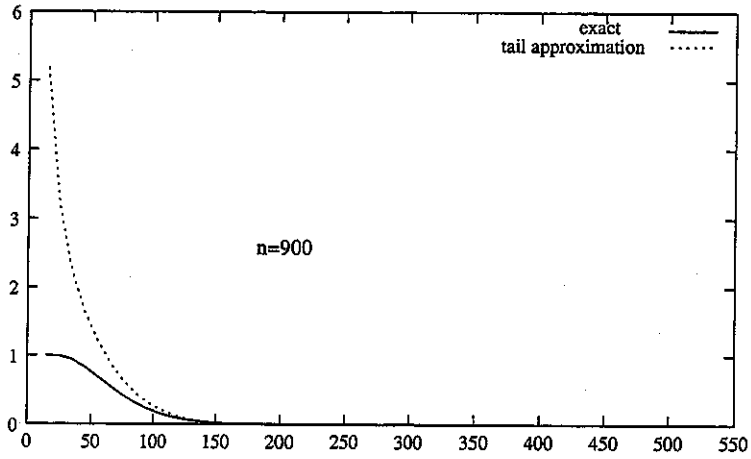

FIG. 5.1 – Approximation de la queue de distribution, $\mathbb{P}(X = 1) = 0.505$, $\mathbb{P}(X = -1) = 0.495$, $E(X) = 0.01$, $n = 900$, $\delta_{900} = 0.3$

5.4 Résultats numériques

Dans cette partie , nous présentons les résultats obtenus en étudiant divers types de suites. Nous avons considéré des suites de variables indépendantes, qui suivent une loi de Bernoulli $\mathcal{B}(p)$. Nous avons fait varier la valeur de p afin d'étudier les différences de comportement selon la valeur de la moyenne et ce pour deux longueurs de suites : 900 et 2500.

5.4.1 La nature des résultats

Plutôt que de présenter les fonctions de répartition obtenues avec chacune des approximations utilisées, nous proposons de présenter les résultats obtenus sous la forme de deux tableaux.

Le premier (tableau 5.2) donne deux critères pour mesurer l'écartement par rapport à la courbe exacte. Nous définissons ainsi les deux quantités Δ et Δ' de la manière suivante : pour le seuil 1% (resp. le seuil 1/1000), Δ (respectivement Δ') mesure l'écart relatif entre l'abcisse de la courbe exacte et l'abcisse de la courbe approchée comme représenté sur la figure 5.3. Le choix des seuils 1% et 1/1000 nous a semblé assez naturel puisque l'on s'intéresse essentiellement aux événements rares.

Le second tableau (5.3, page 85) reprend une part des résulats obtenus dans le premier. Ainsi pour chaque suite étudiée, on garde le score local obtenu avec la distribution exacte et associé au quantile 1/100 (respectivement 1/1000) ; on le note Sexact. Ensuite pour chacune des 3 approximations, on regarde la probabilité d'atteindre un tel score.

Lorsque les valeurs exactes des abcisses correspondant au seuil 1% ou

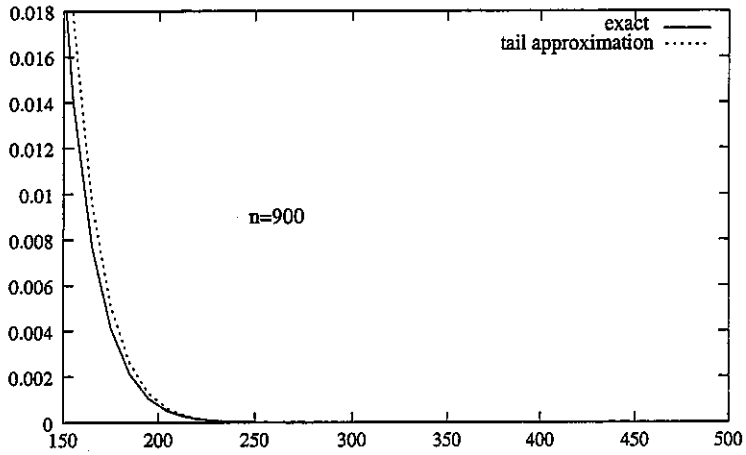

FIG. 5.2 – Approximation de la queue de distribution,Zoom sur la figure 5.1

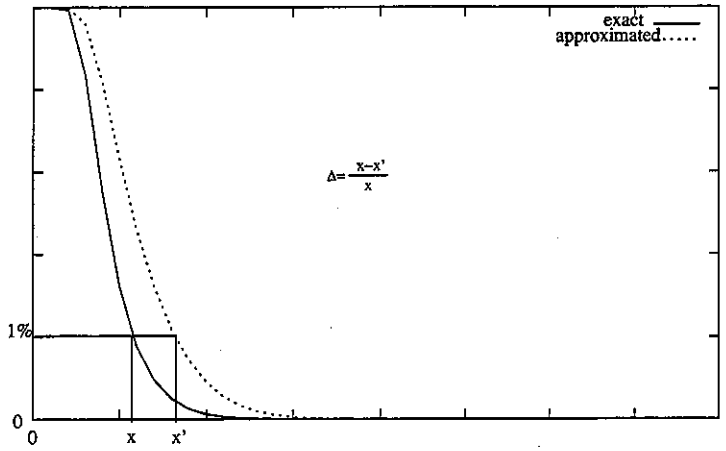

FIG. 5.3 – Que représente Δ ?

1/1000 n'étaient pas disponibles, nous les avons déterminées par interpolation linéaire entre les deux valeurs les plus proches.

5.4.2 Les tableaux récapitulatifs

Cette partie est dédiée aux résultats numériques qui sont présentés sous la forme de deux tableaux. Le premier donne pour chaque approximation, l'erreur relative commise lorsqu'on approche la distribution par une des approximations proposées.

L'indication Nd dans le tableau signifie qu'on ne peut pas déterminer la valeur pour l'approximation considérée. Ce cas se produit le plus fréquemment pour l'approximation de Karlin. En effet, lorsque la moyenne est trop proche de 0, il est très difficile d'atteindre des probabilités de l'ordre de 10^{-2}

ou 10^{-3}. Il est en de même pour l'approximation par la queue de distribution.

Le second tableau donne d'une part la valeur du score local correspondant au quantile 1% (respectivement 1 pour mille) et pour chaque approximation, la probabilité d'observer un tel score.

	Brownien		Karlin		Queue de distribution	
	Δ	Δ'	Δ	Δ'	Δ	Δ'
$\mathbb{E}(X) = 0.002$ $n = 900$	0.011	0.0094	Nd	Nd	0.018	0.011
$\mathbb{E}(X) = 0.002$ $n = 2500$	0.011	0.021	Nd	Nd	0.0088	0.012
$\mathbb{E}(X) = 0$ $n = 900$	0.0055	0.0053	Nd	Nd	0.0054	0.010
$\mathbb{E}(X) = 0$ $n = 2500$	0.0037	0.0028	Nd	Nd	0.0151	0.0083
$\mathbb{E}(X) = -0.002$ $n = 900$	0.024	0.021	2.14	2.54	0.035	0.026
$\mathbb{E}(X) = -0.002$ $n = 2500$	0.033	0.028	Nd	Nd	0.012	0.0067
$\mathbb{E}(X) = -0.02$ $n = 900$	0.21	0.18	0.60	0.82	0.0086	0.0049
$\mathbb{E}(X) = -0.02$ $n = 2500$	0.36	0.31	0.38	0.52	0.0485	0.028
$\mathbb{E}(X) = -0.2$ $n = 900$	2.93	2.86	0.038	0.032	Nd	Nd
$\mathbb{E}(X) = -0.2$ $n = 2500$	4.87	4.86	0.031	0.028	Nd	Nd

Une autre manière de comparer les approximations est proposée dans le tableau suivant. On s'intéresse au quantile 0.01 et on détermine le score correspondant S_{exact} en utilisant la distribution exacte du score local, c'est-à-dire :

$$\mathbb{P}(H_n \leq S_{exact}) = 0.01.$$

On va ensuite regarder quelle est la probabilité d'observer un tel score, lorsque l'on approxime la fonction de répartition du score local par l'uen des trois méthodes étudiées. Si l'approximation est valide, on doit trouver une probabilité proche de 0.01. Les résultats obtenus pour cette comparaison sont présentés dans le tableau 5.3 pour les quantiles 0.01 et 0.01.

			$\mathbb{P}(H_n \geq \text{Sexact})$ avec					
	Sexact		Brownien		Karlin		Tail	
	$Q_{10^{-2}}$	$Q_{10^{-3}}$	$Q_{10^{-2}}$	$Q_{10^{-3}}$	$Q_{10^{-2}}$	$Q_{10^{-3}}$	$Q_{10^{-2}}$	$Q_{10^{-3}}$
$\mathbb{E}(X) = 0.002$ $n = 900$	84	104	1.12e-2	1.05e-3	3.81e-1	3.17e-1	1.36e-2	1.39e-3
$\mathbb{E}(X) = 0.002$ $n = 2500$	142	178	9.94e-3	9.88e-4	1.39	1.39	9.16e-3	8.38e-4
$\mathbb{E}(X) = 0$ $n = 900$	84	104	1.02e-2	1.05e-3	2.71e-1	2.07e-1	1.13e-2	1.13e-3
$\mathbb{E}(X) = 0$ $n = 2500$	140	174	1.02e-2	1.00e-3	7.40e-1	6.63e-1	1.13e-2	1.08e-3
$\mathbb{E}(X) = -0.002$ $n = 900$	82	102	1.25e-2	1.35e-3	2.462e-1	1.85e-1	1.18e-2	1.18e-3
$\mathbb{E}(X) = -0.002$ $n = 2500$	136	170	1.31e-2	1.35e-3	4.58e-1	3.67e-1	1.10e-2	1.03e-3
$\mathbb{E}(X) = -0.02$ $n = 900$	70	88	3.92e-2	6.69e-3	6.57e-2	2.97e-2	9.24e-3	1.06e-3
$\mathbb{E}(X) = -0.02$ $n = 2500$	102	132	8.26e-2	1.66e-2	4.74e-2	1.50e-2	7.68e-3	7.99e-4
$\mathbb{E}(X) = -0.2$ $n = 900$	22	26	8.59e-1	7.37e-1	5.35e-3	1.06e-3	1.18e-10	3.89e-11
$\mathbb{E}(X) = -0.2$ $n = 2500$	24	30	9.93e-1	9.53e-1	6.60e-3	5.79e-4	4.67e-25	1.01e-25

Nous savons que l'approximation pour la queue de distribution n'est valable que pour les valeurs extrêmes. Il n'est donc pas surprenant que son comportement ne soit pas très satisfaisant lorsque l'on s'intéresse aux quantiles 1/100 ou 1/1000. Pour préciser le domaine de validité de cette approximation, regardons des quantiles plus élevés.

Nous avons vu dans la partie 3.4 du chapitre 3 (page 33) qu'il était fréquent en pratique de rencontrer des scores de l'ordre de 10^{-20} voire même de 10^{-40}. Pour des raisons numériques (accès à la valeur exacte à l'aide de la formule (5.2.1)) nous avons uniquement examinés les quantiles 10^{-10} et 10^{-20}. Les scores associés à ces quantiles sont présentés dans les tableaux 5.4 et 5.5.

Le tableau contient :

- le score théorique correspondant au quantile 10^{-10} (resp. 10^{-20}) noté S_{ex},
- la probabilité (P_q) d'observer S_{ex}, probabilité calculée à l'aide de l'équivalent de la queue de distribution
- et enfin le score local (S_q) correspondant au quantile 10^{-10} (resp. 10^{-20}) de l'approximation.

$n = 900$						
	$Q_{10^{-10}}$			$Q_{10^{-20}}$		
	S_{ex}	P_q	S_q	S_{ex}	P_q	S_q
$\mathbb{E}(X) = 0.002$	197	1.6e-10	198	296	3.6e-20	300
$\mathbb{E}(X) = 0$	195	1.6e-10	196	278	3.9e-20	282
$\mathbb{E}(X) = -0.002$	194	1.4e-10	195	277	3.0e-20	280
$\mathbb{E}(X) = -0.02$	179	1.1e-10	179	262	1.2e-20	264
$\mathbb{E}(X) = -0.2$	66	4.4e-16	22	121	6.6e-24	99

$n = 2500$						
	$Q_{10^{-10}}$			$Q_{10^{-20}}$		
	S_{ex}	P_q	S_q	S_{ex}	P_q	S_q
$\mathbb{E}(X) = 0.002$	332	1.2e-10	333	473	1.6e-20	475
$\mathbb{E}(X) = 0$	327	1.3e-10	328	468	1.6-20	470
$\mathbb{E}(X) = -0.002$	322	1.3e-10	323	463	1.6e-20	465
$\mathbb{E}(X) = -0.02$	280	9.8e-11	279	420	1.2e-20	421
$\mathbb{E}(X) = -0.2$	69	5.8e-30	Nd	125	2.8e-36	1

5.5 Analyse des résultats

Les résultats numériques obtenus confirment que l'approximation de Karlin est très mauvaise dans le cas centré, et que l'approximation brownienne est inutilisable lorsque la moyenne est trop élévée.

En fait, le bon critère pour l'approximation brownienne n'est pas la moyenne des variables aléatoires considérées mais $\delta_n = \sqrt{n}\mathbb{E}(X_1)$. Tant que ce critère, en valeur absolue, est inférieur à 0.6 (cas de 900 variables pour $\mathbb{E}(X_1) = -0.02$) l'approximation brownienne donne des résultats tout à fait satisfaisants. Dans tous les cas, elle minore $\mathbb{P}(H_n \geq x)$, ce qui évite de déclarer à tort, que le score local observé est statistiquement significatif. Au contraire, cette approximation aura tendance à masquer des événements exceptionnels.

Lorsque la moyenne est inférieure à -0.2, l'approximation de Karlin donne des très bons résultats, et c'est bien elle qu'il faut utiliser lorsque l'on est dans ce cadre.

Le comportement de l'approximation par la queue de distribution est moins facile à définir. Le tableau 5.4 montre que lorsque l'espérance des variables aléatoires considérées est supérieure à -0.02, on a une très bonne approximation de la queue de distribution du score local en utilisant l'équivalent de la formule (5.2.4). A nouveau le critère important pour cette approximation n'est pas la moyenne mais δ_n. Elle est valable, lorsque l'on s'intèresse à des probabilités de l'ordre de 10^{-10} ou 10^{-20}, tant que $|\delta_n| \leq 1$ (cas de 2500 variables ayant une espérance de -0.02).

5.6 Conclusion

Finalement, il est conseillé d'utiliser
- l'approximation brownienne lorsque $\delta = \sqrt{n}|\mathbb{E}(X_1)| \leq 0.6$,
- l'approximation de Karlin lorsque l'espérance est inférieure à -0.2,
- l'équivalent pour la queue de distribution lorsque $|\delta| \leq 1$ et lorsque l'on s'intéresse à des quantiles au-delà de 10^{-10}.

Le choix entre les différentes approximations n'est pas simple puisqu'il ne dépend pas d'un unique critère. Il faut en effet s'intéresser à la fois à l'espérance et à δ pour pouvoir décider. Nous présentons ici un résumé pour les domaines de validité de chacune des approximations sous la forme de deux graphiques : l'un concerne les suites de longueur 900, l'autre celles de longueur 2500 ; cette présentation permet de ne plus se soucier de δ et de représenter le domaine de validité uniquement en fonction de l'espérance des variables considérées.

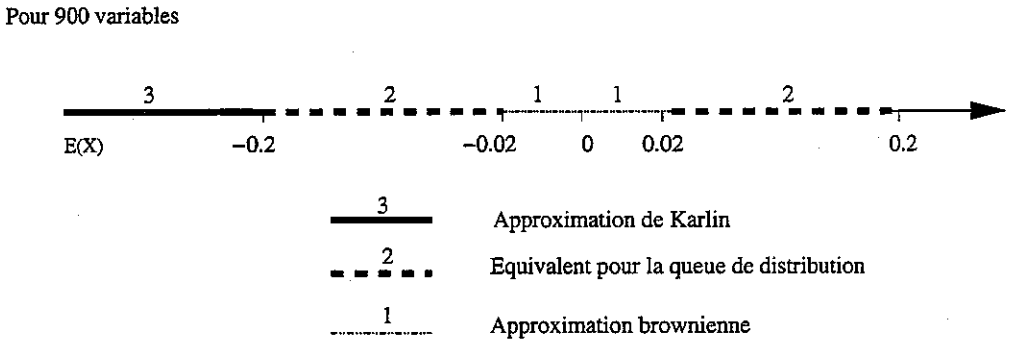

FIG. 5.4 – Domaine de validité dans le cas de 900 variables

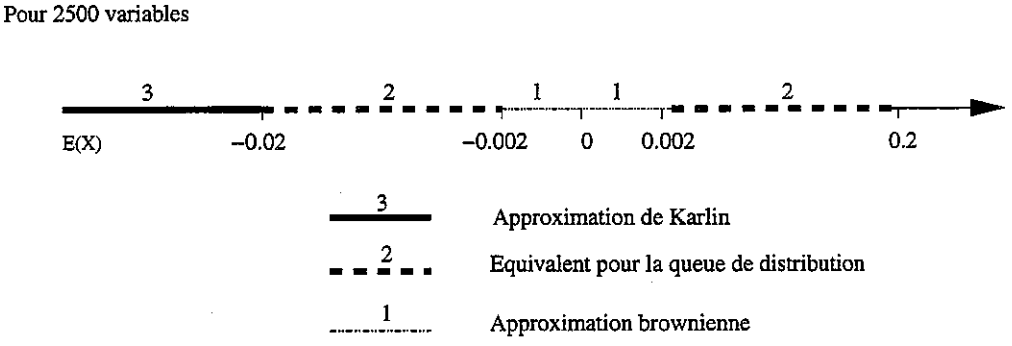

FIG. 5.5 – Domaine de validité dans le cas de 2500 variables

Bibliographie

[BDMR02] 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. *Annals of the Institute of Statistical Mathematics*, December 2002. To appear.

[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.

Chapitre 6

Approximation de la distribution du maximum d'une marche aléatoire. Application au score local.

Contents

Résumé :

Soit $(\xi_i)_{i\geq 1}$ une suite de variables aléatoires i.i.d. bornées par K, d'espérance nulle et de variance σ^2. On note $(X_n)_{n\geq 0}$ la marche aléatoire associée.

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

On sait d'après le chapitre 4 que

$$\frac{H_n}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{(d)} \sup_{0\leq u\leq 1} |B_u|,$$

où $(B_u, u \geq 0)$ désigne un mouvement brownien standard.

Rappelons (cf. chapitre 5) que numériquement, on a l'inégalité suivante :

$$\mathbb{P}\left(B_1^* \geq (x+K)/\sigma\sqrt{n}\right) \leq \mathbb{P}\left(H_n \geq x\right) \leq \mathbb{P}\left(B_1^* \geq x/\sigma\sqrt{n}\right) \quad \forall x \geq 0,$$

où $B_1^* = \sup_{0\leq u\leq 1} |B_u|$.

Cette observation numérique nous conduit naturellement à estimer la vitesse de convergence de $\mathbb{P}\left(H_n/\sqrt{n} \geq x\right)$ vers $\mathbb{P}\left(B_1^* \geq x/\sigma\right)$.

La méthode que nous avons développée utilise un couplage entre une marche aléatoire et un mouvement brownien. Elle s'applique également à d'autres fonctionnelles de $(X_n)_{n\geq 0}$, notamment le maximum d'une marche aléatoire et c'est dans ce cadre que nous présentons l'étude.

Le but du travail présenté dans ce chapitre est donc de donner des bornes effectives à

$$\delta_n(S) = \left| \mathbb{P}\left(\frac{S_n}{\sqrt{n}} \geq x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right) \right|$$

et à

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

On montre (cf. théorème 6.1) l'inégalité suivante :

$$\delta_n(S) \leq \hat{C}(n, K/\sigma)\sqrt{\frac{\ln n}{n}},$$

où

$$\hat{C}(n,y) = \frac{1}{\sqrt{\ln n}}\left(1 + \frac{2y}{\sqrt{2\pi}}\right) + \frac{2e^{-1/2}}{\sqrt{\pi}}\sqrt{\frac{5\,y^4}{3} - 1}.$$

Dans le cas du score local (cf. théorème 6.8), on obtient une identité analogue

$$\delta_n(H) \leq \bar{C}(n, K/\sigma)\sqrt{\frac{\ln n}{n}},$$

où

$$\bar{C}(n,y) = \frac{1}{\sqrt{\ln n}}\left(1 + \frac{4y}{\sqrt{2\pi}}\right) + \frac{2e^{-1/2}}{\sqrt{\pi}}\sqrt{\frac{5\,y^4}{3} - 1}.$$

$\hat{C}(n,y)$ est explicite. Il est toutefois intéressant de tester numériquement la qualité de la borne obtenue. Nous avons donc procédé à des simulations en considérant trois classes de lois pour ξ_i : une classe de distributions uniformes, une autre qui a tendance à charger davantage les bords du support et enfin une troisième classe qui se concentre sur le centre du support.

Ce travail a fait l'objet d'un article accepté dans *Methodology and Computing in Applied Probability*.

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

14th February 2003

Abstract

Let $(X_n)_{n\geq 0}$ be a real random walk starting at 0, with centered increments bounded by a constant K. The main result of this study is : $\left|\mathbb{P}\left(\frac{S_n}{\sqrt{n}} \geq x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq x\right)\right| \leq C(n,K)\sqrt{\frac{\ln n}{n}}$, where $x \geq 0$, σ^2 is the variance of the increments, S_n is the supremum at time n of the random walk, $(B_u, u \geq 0)$ is a standard 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, invariance principle , rate of convergence, local score, maximum.
AMS 2000 Subject classifications
60F05, 60F17, 60G17, 60G40, 60G50, 60J65.

6.1 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 :

$$X_0 = 0, \qquad X_n = \sum_{i=1}^{n} \xi_i, \qquad n \geq 1. \tag{6.1.1}$$

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

$$| \mathbb{P}\left(\frac{X_n}{\sigma\sqrt{n}} \geq x\right) - \mathbb{P}(G \geq x) | \leq \frac{C\mathbb{E}\left[|\xi_1|^3\right]}{\sqrt{n}}; \quad x \in \mathbb{R}, n \geq 1; \qquad (6.1.2)$$

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 :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \geq x\right) = \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq x\right); \quad x \geq 0, \qquad (6.1.3)$$

where $(B_t, t \geq 0)$ is a standard one dimensional Brownian motion started at 0.
Since $\sup_{0\leq u\leq 1} B_u$ and $|B_1|$ are identically distributed, the right hand-side of (6.1.3) can be easily computed.
A priori the rate of convergence of $\mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \geq x\right)$ to $\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq x\right)$ is unknown.
3) In [DEV00], motivated by biological considerations, we established a similar result to (6.1.3) :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{H_n}{\sigma\sqrt{n}} \geq x\right) = \mathbb{P}(B_1^* \geq x); \quad x \geq 0, \qquad (6.1.4)$$

where $H_n = \max_{0\leq i\leq j} (X_j - X_i)$ 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 $n \to |\mathbb{P}\left(\frac{H_n}{\sigma\sqrt{n}} \geq x\right) - \mathbb{P}(B_1^* \geq x)|$ is unknown. Therefore its knowledge would be useful.

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

$$\delta_n(S) = | \mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \geq x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq x\right) |$$

and to

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

We prove (cf theorems 6.1 and 6.8) 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 6.2 we deal with the supremum of a centred random walk. The previous analysis can be adapted (cf section 6.3) to handle the local score and is presented in section 6.3. In section 6.3.2, with the process S, we check the accuracy of previous bounds through numerical tests.

6.2 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

$$X_0 = 0, \quad X_n = \sum_{i=1}^{n} \xi_i, \quad n \geq 1. \tag{6.2.1}$$

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

$$|\xi_n| \leq K, \quad \forall n \geq 1. \tag{6.2.2}$$

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

$$\text{the law of } B_T \text{ is } \mu, \tag{6.2.3}$$

and

$$(B_{T\wedge t}, t \geq 0) \text{ is a uniformly integrable martingale.} \tag{6.2.4}$$

(6.2.4) 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 (6.2.4) imply :

$$T \leq T^*(A), \tag{6.2.5}$$

where $T^*(A) = \inf\{t \geq 0, |B_t| \geq A\}$.
Conversely (6.2.5) implies (6.2.4).
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 (6.2.5).
Let $\mathcal{P}_c$ be 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 :

$$B_{U(\mu)} \sim \mu, \quad Supp(\mu) \subset [-K, K], \quad U(\mu) \leq T^*(K). \tag{6.2.6}$$

In particular if μ belongs to $\mathcal{P}_c$, we have the useful identity :

$$\mathbb{E}\left[\left(B_{U(\mu)}\right)^2\right] = \mathbb{E}\left[U(\mu)\right] < +\infty. \tag{6.2.7}$$

We need a little bit more than (6.2.6), we assume $\mathcal{P}_c$ has the following scaling property :

$$U(\mu_c) \overset{(d)}{=} c^2 U(\mu), \text{ for any } c > 0, \tag{6.2.8}$$

where μ_c is the image of μ by $x \mapsto cx$.
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 6.1 For all $x \geq 0$ and $M \geq 2$:

$$\left|\mathbb{P}\left(\frac{S_M}{\sqrt{M}} \geq x\right) - \mathbb{P}\left(\sup_{0 \leq u \leq 1} B_u \geq \frac{x}{\sigma}\right)\right| \leq C(M, \mu)\sqrt{\frac{\ln M}{M}}. \tag{6.2.9}$$

where μ is the common distribution of ξ_i and

$$C(M, \mu) = \frac{2K}{\sigma\sqrt{2\pi}} \frac{1}{\sqrt{\ln M}} + \frac{1}{\sqrt{\ln M}} + \frac{2\, e^{-1/2} \sqrt{\mathbb{E}(U(\mu)^2) - \sigma^4}}{\sqrt{\pi}\sigma^2}. \tag{6.2.10}$$

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

$$\hat{C}(M, y) = \frac{1}{\sqrt{\ln M}}\left(1 + \frac{2}{\sqrt{2\pi}} y\right) + \frac{2\, e^{-1/2}}{\sqrt{\pi}} \sqrt{\frac{5y^4}{3} - 1}. \tag{6.2.11}$$

The upper bound $C(M, \mu)$ is the best possible, given by our approach. Obviously we need the value of $\mathbb{E}\left[U(\mu)^2\right]$. We notice that $\hat{C}(M, K/\sigma)$ is easy to determine since it only depends on the two key parameters K and σ. For

instance let α be a positive parameter belonging to $]0;1[$, μ be a centred probability measure with compact support in $[-K;K]$ and σ^2 its variance. Since $\hat{C}(M,K/\sigma)\sqrt{\ln M/M}$ goes to 0, as M goes to infinity, we can choose M such that $\hat{C}(M,K/\sigma)\sqrt{\ln M/M} \leq \alpha$.
Consequently (6.2.9) implies that

$$|\mathbb{P}\left(\frac{S_M}{\sqrt{M}} \geq x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right)| \leq \alpha.$$

3) 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 6.2 There exists a sequence of stopping times $(T_n)_{n\geq 0}$, such that :

$$T_0 = 0, \quad T_k = \sum_{1\leq i\leq k} T_i', \tag{6.2.12}$$

and

$$(\sigma B_{T_k}, k \geq 0) \stackrel{(d)}{=} \left(\frac{X_k}{\sqrt{M}}, k \geq 0\right), \tag{6.2.13}$$

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 (6.2.6) 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 :

$$T_1' = T_1, \tag{6.2.14}$$

$$B_{T_k+T_{k+1}'} - B_{T_k} = B_{T_{k+1}} - B_{T_k} \stackrel{(d)}{=} \frac{1}{\sigma\sqrt{M}}\xi_{k+1}; \quad \forall k \geq 0, \tag{6.2.15}$$

where

$$T_0 = 0, \quad T_k = T_1' + \ldots + T_k'; \quad k \geq 1. \tag{6.2.16}$$

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

$$(B_{T_k} - B_{T_{k-1}}; k \geq 1) \stackrel{(d)}{=} \left(\frac{\xi_k}{\sigma\sqrt{M}}; k \geq 1\right). \tag{6.2.17}$$

□

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 6.2, we take :

$$X_k = \sigma\sqrt{M}B_{T_k}, \quad \forall\, k \geq 1. \tag{6.2.18}$$

We use the strength of (6.2.18) 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 6.3 We have :

$$\frac{1}{\sqrt{M}}S_k \leq \sigma \sup_{0\leq u\leq T_k} B_u \leq \frac{1}{\sqrt{M}}S_k + \frac{K}{\sqrt{M}}; \qquad \forall k \geq 1, \tag{6.2.19}$$

$$\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{1}{\sigma\sqrt{1-\varepsilon}}\left(x + \frac{K}{\sqrt{M}}\right)\right) - \mathbb{P}\left(|T_M - 1| \geq \varepsilon\right) \leq \mathbb{P}\left(\frac{S_M}{\sqrt{M}} \geq x\right), \tag{6.2.20}$$

$$\mathbb{P}\left(\frac{S_M}{\sqrt{M}} \geq x\right) \leq \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma\sqrt{1+\varepsilon}}\right) + \mathbb{P}\left(|T_M - 1| \geq \varepsilon\right), \tag{6.2.21}$$

for any $x \geq 0$ and $\varepsilon > 0$.

Proof : a) (6.2.18) implies (6.2.19).
b) Let $\varepsilon > 0$ and $x \geq 0$. The first inequality in (6.2.19) implies :

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

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

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

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

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

for any $c > 0$,

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

This achieves the proof of (6.2.21).
c) (6.2.20) is a direct consequence of the following inclusions :

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

□

We note that (6.2.16) and (6.2.8) imply that

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

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 :

$$\mathbb{E}[\exp\{\lambda U(\mu)\}] < +\infty \Leftrightarrow \lambda < A(\mu). \tag{6.2.22}$$

Lemma 6.4 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 :

$$\mathbb{P}(T_M - 1 \geq \varepsilon) \leq \exp\{-M f_\varepsilon(\lambda_1)\}, \tag{6.2.23}$$

$$\mathbb{P}(T_M - 1 \leq -\varepsilon) \leq \exp\{-M g_\varepsilon(\lambda_2)\}, \tag{6.2.24}$$

where

$$f_\varepsilon(x) = \sigma^2(1+\varepsilon)x - \ln\left(\mathbb{E}\left[\exp\left(xU(\mu)\right)\right]\right), \quad x < A(\mu), \tag{6.2.25}$$

and

$$g_\varepsilon(x) = -\sigma^2(1-\varepsilon)x - \ln\left(\mathbb{E}\left[\exp\left(-xU(\mu)\right)\right]\right), \quad x \geq 0. \tag{6.2.26}$$

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

$$\begin{aligned} \mathbb{P}\left(T_M \geq 1+\varepsilon\right) &= \mathbb{P}\left(\exp\left\{\lambda\left(T_1' + \ldots + T_M'\right)\right\} \geq \exp\left\{\lambda(1+\varepsilon)\right\}\right) \\ &\leq e^{-\lambda(1+\varepsilon)}\left(\mathbb{E}\left[e^{\lambda T_1}\right]\right)^M. \end{aligned} \tag{6.2.27}$$

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 (6.2.8) :

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

Then

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

(6.2.23) follows immediately.

2) As for (6.2.24) it is sufficient to replace (6.2.27) 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 6.5 There exists $0 < A' \leq A(\mu)$ such that for any $\varepsilon \leq \frac{\mathbb{E}[U(\mu)^2]-\sigma^4}{\sigma^2}A'$,

$$\mathbb{P}\left(|T_M - 1| \geq \varepsilon\right) \leq 2\exp\left(-c_1(\mu)M\varepsilon^2\right), \tag{6.2.28}$$

where

$$c_1(\mu) = \frac{\sigma^4}{4\left(\mathbb{E}(U(\mu)^2) - \sigma^4\right)} > 0. \tag{6.2.29}$$

Proof : 1) According to lemma 6.4, the search of an upper bound for $\mathbb{P}(T_M - 1 \geq \varepsilon)$ leads us to study f_ε. At this step, ε and μ are fixed, f stands for f_ε and $U(\mu)$ will be designated by U.
2) We have

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

Then

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

and

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

where

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

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

$$f(x) \geq h(x) - \frac{\mathbb{E}[U^2] - \sigma^4}{4}x^2, \qquad \forall x \in [0, A']. \tag{6.2.30}$$

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

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

Using (6.2.30), we obtain :

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

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

$$\mathbb{P}(T_M - 1 \geq \varepsilon) \leq \exp\{-c_1(\mu)M\varepsilon\}, \quad \forall \varepsilon \leq \frac{\mathbb{E}[U^2] - \sigma^4}{\sigma^2}A', \tag{6.2.31}$$

where

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

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

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

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

By the same way as previously, we have

$$\mathbb{P}(T_M - 1 \leq -\varepsilon) \leq \exp\{-c_1(\mu)M\varepsilon\}, \quad \forall \varepsilon \leq \frac{\mathbb{E}[U^2] - \sigma^4}{\sigma^2}A'. \quad (6.2.32)$$

□

Lemma 6.6 For any $0 < \varepsilon < 1/2$,

$$\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma\sqrt{1+\varepsilon}}\right) \leq c\varepsilon + \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right), \quad (6.2.33)$$

where

$$c = \frac{1}{2}\sqrt{\frac{3}{\pi}}e^{-1/2}.$$

Proof : As $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma\sqrt{1+\varepsilon}}\right) = \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right) + \delta,$$

where

$$\delta = \mathbb{P}\left(\frac{x}{\sigma\sqrt{1+\varepsilon}} \leq \sup_{0\leq u\leq 1} B_u \leq \frac{x}{\sigma}\right).$$

But it is well known that $\sup_{0\leq u\leq 1} B_u \overset{(d)}{=} |B_1|$, so that :

$$\begin{aligned}\delta &= \mathbb{P}\left(\frac{x}{\sigma\sqrt{1+\varepsilon}} \leq |B_1| \leq \frac{x}{\sigma}\right) = 2\,\mathbb{P}\left(\frac{x}{\sigma\sqrt{1+\varepsilon}} \leq B_1 \leq \frac{x}{\sigma}\right) \\ &= 2\left(\Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma\sqrt{1+\varepsilon}}\right)\right),\end{aligned}$$

with

$$\Phi(z) = \mathbb{P}(B_1 \leq z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-u^2/2}\mathrm{d}u.$$

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

$$0 < \frac{x}{\sigma} - \frac{x}{\sigma\sqrt{1+\varepsilon}} = \frac{x\varepsilon}{\sigma\sqrt{1+\varepsilon}\left(\sqrt{1+\varepsilon}+1\right)} \leq \frac{x\varepsilon}{2\sigma}.$$

Suppose that $\varepsilon < 1/2$ and $y \in \left[\frac{x}{\sigma\sqrt{1+\varepsilon}}, \frac{x}{\sigma}\right]$, then

$$\Phi'(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \leq \frac{1}{\sqrt{2\pi}} e^{-x^2/(3\sigma^2)}.$$

So that

$$\delta \leq \varepsilon h_0\left(\frac{x}{\sigma}\right),$$

where

$$h_0(z) = \frac{z}{\sqrt{2\pi}} e^{-z^2/3}.$$

But $h_0(z) \leq h_0(\sqrt{3/2}) = c$, this shows (6.2.33). □

At this stage we have to give a lower bound to $\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{1}{\sigma\sqrt{1-\varepsilon}}\left(x + \frac{K}{\sigma\sqrt{M}}\right)\right)$. Using same tools as for lemma 6.6, we will prove :

Lemma 6.7 For any $0 < \varepsilon < 1/2$,

$$\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right) - \frac{2K}{\sigma\sqrt{2\pi M}} - c_2\,\varepsilon \leq \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{1}{\sigma\sqrt{1-\varepsilon}}\left(x + \frac{K}{\sqrt{M}}\right)\right) \quad (6.2.34)$$

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 6.6, 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 $\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq x/\sigma\right)$.

$$\begin{aligned}
\mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq x/\sigma\right) &- \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq y/\sigma\right) \\
&= \mathbb{P}\left(x/\sigma \leq \sup_{0\leq u\leq 1} B_u \leq x/\sigma + K/(\sigma\sqrt{M})\right), \\
&= 2\left(\Phi\left(\frac{x}{\sigma} + \frac{K}{\sigma\sqrt{M}}\right) - \Phi\left(\frac{x}{\sigma}\right)\right), \\
&\leq \frac{2K}{\sigma\sqrt{2\pi M}}\, e^{-x^2/(2\sigma^2)}, \\
&\leq \frac{2K}{\sigma\sqrt{2\pi M}}.
\end{aligned}$$

This ends the proof. □

We are now able to prove theorem 6.1. We can control the rate of convergence of the two probability distributions functions.

Proof of theorem 6.1 : Using lemma 6.3, (6.2.28), (6.2.33) and (6.2.34), we obtain :

$$\begin{aligned}
\left|\mathbb{P}\left(\frac{S_M}{\sqrt{M}} \geq x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right)\right| & \\
\leq \max\Bigg\{ & \frac{2K}{\sigma\sqrt{2\pi M}} + \frac{2e^{-1/2}}{\sqrt{2\pi}}\varepsilon + 2\exp\left\{-c_1(\mu)M\varepsilon^2\right\}, \\
& \frac{1}{2}\sqrt{\frac{3}{\pi}} e^{-1/2}\varepsilon + 2\exp\left\{-c_1(\mu)M\varepsilon^2\right\}\Bigg\}.
\end{aligned}$$

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 :

$$\begin{aligned} &|\mathbb{P}\left(\frac{S_M}{\sqrt{M}} \geq x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right)| \\ &\qquad \leq \frac{2K}{\sigma\sqrt{2\pi M}} + \frac{2e^{-1/2}}{\sqrt{2\pi}}\sqrt{\frac{\ln M}{2\,M\,c_1(\mu)}} + \frac{1}{\sqrt{M}}, \\ &\qquad \leq \sqrt{\frac{\ln M}{M}}\left(\frac{2K}{\sqrt{\ln M}\sigma\sqrt{2\pi}} + \frac{1}{\sqrt{\ln M}} + \frac{e^{-1/2}}{\sqrt{\pi\,c_1(\mu)}}\right). \end{aligned}$$

Since $c_1(\mu)$ is given by (6.2.29), (6.2.9) 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

$$\mathbb{E}\left[U(\mu)^2\right] \leq \mathbb{E}\left[(T^*(K))^2\right] = K^4\mathbb{E}\left[(T^*(1))^2\right] = \frac{5}{3}K^4.$$

□

6.3 Applications to the local score. Numerical tests.

If we replace $(X_n)_{n\geq 0}$ by $(-X_n)_{n\geq 0}$ in theorem 6.1 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 :

$$|\mathbb{P}\left(\frac{\min_{0\leq i\leq M} X_i}{\sqrt{M}} \leq -x\right) - \mathbb{P}\left(\sup_{0\leq u\leq 1} B_u \geq \frac{x}{\sigma}\right)| \leq \hat{C}(M, K/\sigma)\sqrt{\frac{\ln M}{M}},$$

$$\hat{C}(M,y) = \frac{1}{\sqrt{\ln M}}\left(1 + \frac{2}{\sqrt{2\pi}}y\right) + \frac{2\,e^{-1/2}}{\sqrt{\pi}}\sqrt{\frac{5y^4}{3} - 1}.$$

Our scheme developed previously in section 6.2 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 6.1 for the local score (theorem 6.8).
We also end up this paper with numerical computations.

6.3.1 The local score

As we did in section 6.2, 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 (6.1.1)). The local score H_M of $(X_n)_{n\geq 0}$ is

$$H_M = \max_{0\leq i\leq j\leq M}(X_j - X_i) = \max_{0\leq j\leq M}\left(X_j - \min_{0\leq i\leq j} X_i\right). \tag{6.3.1}$$

Let us state our main result involving the local score.

Theorem 6.8 For all $x \geq 0$, $M \geq 2$,

$$|\mathbb{P}\left(\frac{H_M}{\sqrt{M}} \geq x\right) - \mathbb{P}\left(\sigma \sup_{0\leq u\leq 1}|B_u| \geq x\right)| \leq \bar{C}(M, K/\sigma)\sqrt{\frac{\ln M}{M}}, \tag{6.3.2}$$

where

$$\bar{C}(M,y) = \frac{1}{\sqrt{\ln M}}\left(1 + \frac{4}{\sqrt{2\pi}}y\right) + \frac{2\,e^{-1/2}}{\sqrt{\pi}}\sqrt{\frac{5y^4}{3} - 1}.$$

We are now able to prove the main result about the behaviour of the local score.

Proof of theorem 6.8 : The method is the same as the one developed for the maximum. However there are two changes.
a) (6.2.19) 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. □

6.3.2 Numerical tests

This section is devoted to the numerical validation of our results : we would like to verify the quality of our upper bound $\hat{C}(M, K/\sigma)$ in (6.2.11).

Three classes of examples of μ.

For simplicity we consider only discrete probability measures. Let us recall that μ is the common distribution of ξ_i.
We examine three classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ of μ.

- $\mathcal{C}_1$ is constituted by uniform distributions on $\{-2,2\}$, $\{-5,\ldots,-1,1,\ldots,5\}$ and $\{-10,\ldots,-1,1,\ldots,10\}$, noted $\mu_{1,1}$, $\mu_{1,2}$ and $\mu_{1,3}$ respectively.

- In $\mathcal{C}_2$ the three probability measures $\mu_{2,1}$, $\mu_{2,2}$ and $\mu_{2,3}$ are rather concentrated at the end points of their support. More precisely we choose :

$$\mu_{2,1} = \frac{1}{6} \sum_{i=-2, i\neq 0}^{2} |i| \delta_i,$$

$$\mu_{2,2} = \frac{1}{30} \sum_{i=-5, i\neq 0}^{5} |i| \delta_i,$$

$$\mu_{2,3} = \frac{1}{110} \sum_{i=-10, i\neq 0}^{10} |i| \delta_i,$$

  where δ_i denotes the Dirac measure at i.

- In $\mathcal{C}_3$ we consider $\mu_{3,1}$, $\mu_{3,2}$ and $\mu_{3,3}$ which are rather concentrated at the origin. We take :

$$\mu_{3,1} = \frac{1}{6} \sum_{i=-2, i\neq 0}^{2} (3 - |i|) \delta_i,$$

$$\mu_{3,2} = \frac{1}{30} \sum_{i=-5, i\neq 0}^{5} (6 - |i|) \delta_i,$$

$$\mu_{3,3} = \frac{1}{110} \sum_{i=-10, i\neq 0}^{10} (11 - |i|) \delta_i.$$

We observe that $K = 2$ (resp. $K = 5$, $K = 10$) for $\mu_{i,1}$ (resp. $\mu_{i,2}$, $\mu_{i,3}$), $1 \leq i \leq 3$.

Procedure and results

Let us explain our numerical procedure. Let us start with M fixed. We generate k times the random walk $(X_i)_{0\leq i\leq M}$ and then obtain a k-sample of $S_M/\sqrt{M}$ whose empirical distribution function is denoted $F_{k,M}$.

On one hand, Theorem 6.1 tells us

$$\sup_{x\in\mathbb{R}} \mid \mathbb{P}\left(\frac{S_M}{\sqrt{M}} \leq x\right) - F\left(\frac{x}{\sigma}\right) \mid \leq \hat{C}(M, K/\sigma)\sqrt{\frac{\ln M}{M}} \tag{6.3.3}$$

where $F(x) = \mathbb{P}(|B_1| \leq x)$ and

$$\hat{C}(M, y) = \frac{1}{\sqrt{\ln M}}\left(1 + \frac{2y}{\sqrt{2\pi}}\right) + \frac{2e^{-1/2}}{\sqrt{\pi}}\sqrt{\frac{5\,y^4}{3} - 1}.$$

On the other hand,

$$\sup_{x\in\mathbb{R}} \mid F_{k,M}(x) - F\left(\frac{x}{\sigma}\right) \mid \leq \delta_{k,M}\sqrt{\frac{\ln M}{M}} \tag{6.3.4}$$

with

$$\delta_{k,M} = \sqrt{\frac{M}{\ln M}}\left(\sup_{x\in\mathbb{R}} |F_{k,M}(x) - F(x/\sigma)|\right). \tag{6.3.5}$$

Kolmogorov's theorem implies that $\mathbb{P}\left(S_M/\sqrt{M} \leq x\right)$ can be approximated by $F_{k,M}(x)$, uniformly with respect to x, with heuristic rate $1/\sqrt{k}$. We choose $k = 10^6$.

This brings us to compare $\hat{C}(M, K/\sigma)$ and $\delta_{k,M}$. We introduce

$$R(M, K) = \frac{\hat{C}(M, K/\sigma)}{\delta_{k,M}}. \tag{6.3.6}$$

Then $R(M, K)$ close to 1 (resp. large) means that our upper bound $\hat{C}(M, K/\sigma)$ is convenient (resp. over-estimated).

We plot $M \mapsto R(M, K)$ from $M = 10$ to $M = 1000$, for $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$. This leads to figures 6.1, 6.2 and 6.3 on pages 109 to 110.

To complete this comparison we give the maximum and the minimum over M of $\delta_{k,M}$ and $R(M, k)$, μ belonging to $\mathcal{C}_1$, $\mathcal{C}_2$ or $\mathcal{C}_3$ (see table 6.1).

Class	K	$\delta_{k,M}$		$R(M,K)$	
		min	max	min	max
1	2	0.3	0.5	8.5	12
	5	0.23	0.4	33	52
	10	0.22	0.39	66	110
2	2	0.26	0.44	14	23
	5	0.23	0.4	16	25
	10	0.21	0.38	15	25
3	2	0.25	0.45	17	30
	5	0.24	0.4	26	42
	10	0.22	0.39	32	52

Conclusions about numerical results

Since the measures in $\mathcal{C}_2$ are concentrated at the end points of $[-K, K]$, it seems natural that $\hat{C}(M, K/\sigma)$ is convenient. Numerical simulations confirm this fact.
As for $\mathcal{C}_1$ and $\mathcal{C}_3$ we observe that $\hat{C}(M, K/\sigma)$ is over-estimated and the worse case is realized with uniform distributions (i.e. the class $\mathcal{C}_1$).

But we have to keep in mind that the rate of convergence of $\mathbb{P}\left(\frac{S_M}{\sqrt{M}} \leq x\right)$ to $\mathbb{P}\left(|B_1| \leq x/\sigma\right)$ is $\hat{C}(M, K/\sigma)\sqrt{\frac{\ln M}{M}}$. It would be interesting to take M larger than 1000. However we have been limited by the time of computation. For instance the simulation of $k \times M = 10^6 \times 10^3 = 10^9$ ξ_i r.v.'s (with common distribution $\mu_{1,1}$) and the determination of the maximum take about seven hours, on a bi-processor PowerEdge 2400 under linux Red Hat 6.2 equipped with 1Go of RAM.

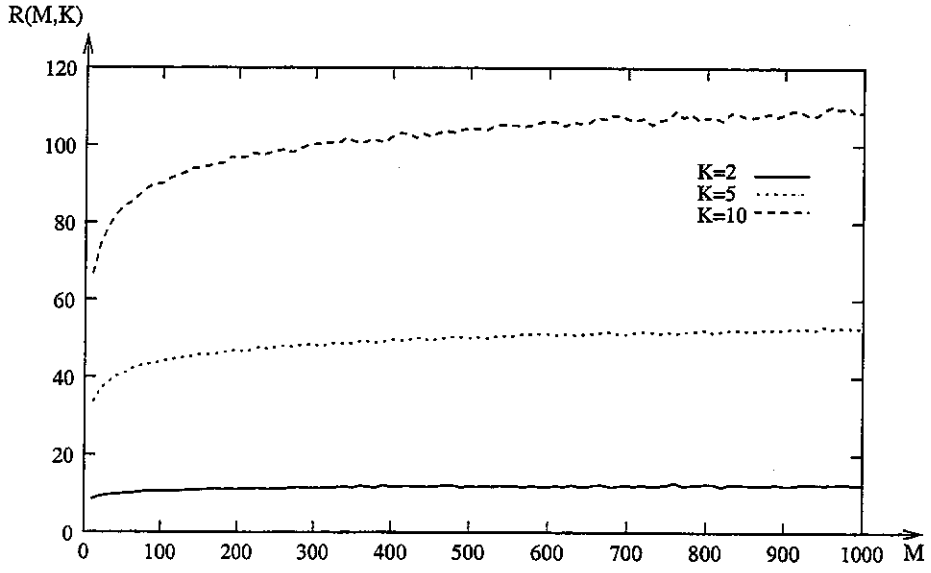

Figure 6.1: Behaviour of the ratio $R(M, K)$ for the class 1

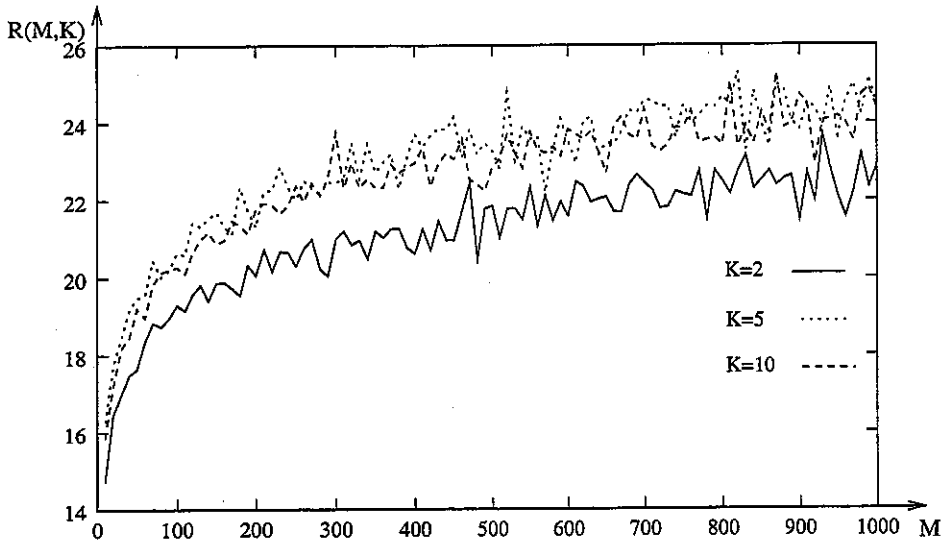

Figure 6.2: Behaviour of the ratio $R(M, K)$ for the class 2

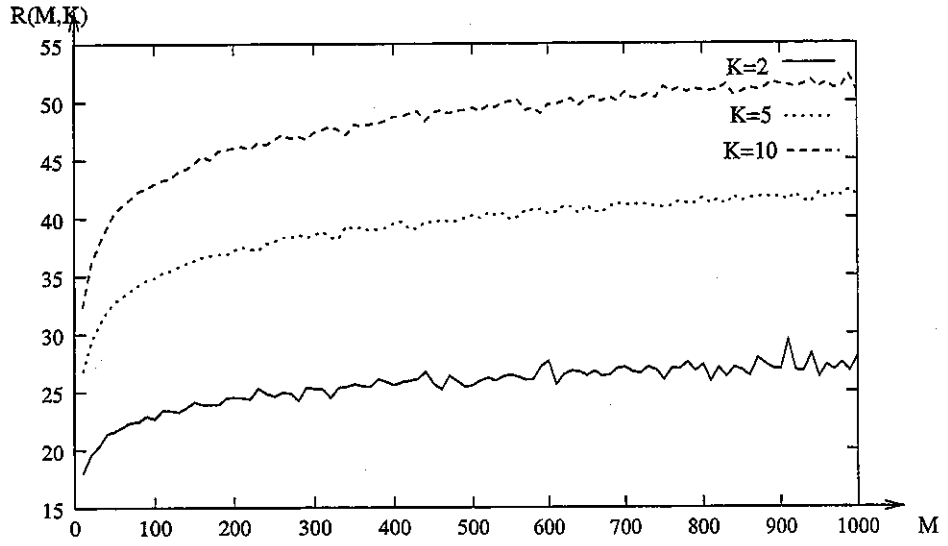

Figure 6.3: Behaviour of the ratio $R(M, K)$ for the class 3

Bibliography

[AY79] J. Azema and M. Yor. Une solution simple au problème de Skorokhod. *Séminaire de probabilités XIII, Univ. Strasbourg 1977/78, Lect. Notes Math.*, 721:90–115, 1979.

[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. *Prépublication de l'Institut Elie Cartan*, 2000. Submitted to Stochastic Processes and their Applications.

[Loè79] M. Loève. *Probability theory.* Springer-Verlag, New York, fourth edition, 1979. Graduate Texts in Mathematics, Vol. 46.

[Sko65] A.V. Skorokhod. *Studies in the theory of random processes.* Reading, Mass.: Addison-Wesley Publish. Comp. Inc. VIII, 1965.

[Val83] P. Vallois. Le problème de Skorokhod sur R: Une approche avec le temps local. *Lect. Notes Math.*, 986:227–239, 1983. Sémin. de probabilités XVII, Proc. 1981/82,.

Université HENRI POINCARÉ
NANCY 1

Mademoiselle **ETIENNE Marie-Pierre**

DOCTORAT de l'UNIVERSITE HENRI POINCARE, NANCY 1

en MATHEMATIQUES APPLIQUEES

VU, APPROUVÉ ET PERMIS D'IMPRIMER

n°783

Nancy, le 16 janvier 2003

Le Président de l'Université

UNIVERSITÉ HENRI POINCARÉ NANCY 1 ★ LE PRÉSIDENT

Cl. BURLET

Université Henri Poincaré, Nancy I
24-30 rue Lionnois - B.P. 3069 - 54013 NANCY Cédex
Tél. : 03 83 68 20 00 - Fax : 03 83 68 21 00

Le score local : un outil pour l'analyse de séquences biologiques.

Les molécules biologiques que sont l'ADN, les différents ARN et les protéines sont à la base des mécanismes du vivant. On peut les considérer comme de longues séquences écrites à l'aide d'un alphabet $\mathcal{A}$ fini. Une des méthodes pour analyser l'information contenue dans ces séquences consiste à attribuer un poids appelé score à chaque composant élémentaire. Le score global de la séquence est alors la somme des scores élémentaires et le **score local** est le maximum des scores de toutes les sous séquences. Le problème statistique qui se pose est d'évaluer le niveau de significativité du score local obtenu.

Pour répondre à cette question, on se place sous l'hypothèse nulle H_0 qui correspond dans cette étude à considérer les scores élémentaires comme des variables i.i.d. Selon le signe de l'espérance des scores élémentaires, le comportement du score local est totalement différent. On détermine alors la loi asymptotique du score local dans le cas centré. Cette étude complète l'approche proposée par Karlin & *al.* et Mercier (espérance négative).

Il existe donc une zone de transition de phase lorsque l'espérance est proche de 0 et nous étudions le comportement numérique des différentes approximations dans ce cas.

Finalement, nous donnons la vitesse de convergence de la fonction de répartition du score local sur une séquence de longueur n lorsque n tend vers l'infini.

Mots clés : score local, score de séquences biologiques, mouvement brownien, P-value, marches aléatoires, théorèmes limites, significativité.

The local score : a tool for the analysis of biological sequences.

For any organism, DNA, RNA and proteins information can be considered as long sequences of letters taken from a finite alphabet $\mathcal{A}$. One way to analyze this information is to assign a weight at each letter (an elementary score). Then, we make the sum over each possible segment and search the segment which realizes the maximal score called **the local score**. Then the problem is to give a level of significance for this local score.

We are led to study the distribution of the local score under the null hypothesis : elementary scores are i.i.d. random variables. According to the sign of the mean, the behaviour of the local score is widely different. We determine the asymptotic distribution of the local score when random variables are centered. This work completes the asymptotic study of Karlin & *al.* and Mercier (the negative expectation).

In the area of phase transition we study numerically the behaviour of the different approximations.

Finally we give the rate of convergence of the cumulative distribution function for the local score over a sequence of length n, as n goes to infinity.

Key-words : local score, biological sequences score Brownian motion, P-value, random walk, limit theorem, significativity.