Markov Controlled Excursions, Local Alignment and Structure

From Markov additive processes to biological sequence analysis

Ph.D. Thesis

Niels Richard Hansen
Department of Applied
Mathematics and Statistics
University of Copenhagen
November 2003

Thesis Supervisors:	Prof. Michael Sørensen, University of Copenhagen
	Associate Prof. Ernst Hansen, University of Copenhagen
Thesis Committee:	Prof. Thomas Mikosch, University of Copenhagen
	Prof. Jens Ledet Jensen, University of Aarhus
	Prof. Ola Hössjer, Stockholm University

Preface

This thesis constitutes my Ph.D.-thesis – a part of the requirements for achieving the Ph.D. degree at the Faculty of Sciences, University of Copenhagen. The thesis work was carried out from November 2000 to October 2003 at the Department of Applied Mathematics and Statistics, University of Copenhagen, funded by a grant from the University of Copenhagen.

The subject of the thesis is probability and bioinformatics – as viewed from a person inclined to do mathematical research. As it turned out, the thesis is highly mathematical and the scientific contribution of the thesis is mostly to the theory of probability and mathematical statistics. Theories and results about the tools used in bioinformatics – in particular biological sequence analysis – are developed, and hopefully this will result in improved tools for other bioinformaticians to benefit from.

The subject of bioinformatics and biological sequence analysis was kindly suggested by my supervisors, Professor Michael Sørensen and Associate Professor Ernst Hansen, whom I owe a lot for their qualified guidance throughout this project. They have made up a good team. Ernst and Michael have also played a key role in establishing good and useful connections to biologist and bioinformaticians. I will mention Professor Peter Arctander and his group at the Department of Evolutionary Biology, University of Copenhagen, together with Professor Anders Krogh and the Bioinformatics Centre, University of Copenhagen. They have provided me with much inspiration for the practical applications of the work presented in this thesis. In this connection I want to thank Morten Lindow at the Bioinformatics Centre, who extracted the RNA-data presented in Chapter 7 in a useful format. I would also like to thank Professor Olle Neerman, Chalmers University of Technology, and Professor David Siegmund, Stanford University, for hosting me when visiting Chalmers and Stanford respectively. In particular, I would like to thank David Siegmund for useful

suggestions in relation to an early version of Chapter 5 and Chapter 6 and for many inspiring discussions about statistics and probability theory in biological sequence analysis.

Furthermore, I would like to thank everybody at the Department of Applied Mathematics and Statistics for all having contributed to this thesis in one way or the other. I want to mention Martin Jacobsen with whom I have had some valuable discussions, and, most notably, I want to thank Anders Tolver Jensen for not just being a good colleague and a close collaborator, with whom I have discussed my work for many hours, but also for being a very good friend. Our collaboration was fruitful and the results presented in Chapter 3 of this thesis are jointly produced. I also want to thank Anders Tolver Jensen for help with proofreading part of this manuscript.

Last but not least I want to thank my small but enlarging family, Marianne and little Mads, for their love and support – Mads for mostly being. Without them all this would matter very little.

The publication status of the new results achieved in the thesis is as follows: The results in Chapter 3, in a slightly different form, have appeared in the preprint Hansen & Jensen (2003), and this paper has been accepted for publications in *Stochastic Processes and their Applications*. The results presented in Chapter 5 and 6 are expected to be submitted shortly as two papers – hopefully to be published consecutively in the same journal. Finally the more practical work presented in Chapter 7 will most likely be submitted in an extended version including some work on the ideas discussed at the end of that chapter.

In spite of the assistance, help and guidance I have received, I am in the end responsible for the content of the thesis. Thus any remaining errors or misprints are entirely my fault.

November 2003
Niels Richard Hansen

Summary

In this thesis we study the comparison of sequences from a finite alphabet and the theory of excursions for Markov additive processes. The main motivation is the applications to biological sequence analysis. A central problem is to test if two sequences violate the hypothesis of being *independent* random sequences from a specific model with a test statistic chosen to reflect the specific kind of violation that we are interested in. Another related problem introduced in this thesis is to test if one random sequence contains certain structural features. Considering *local* violations of the hypothesis when comparing independent sequences of *iid* variables, results have been obtained by e.g. Dembo et al. (1994*b*) based on the behaviour of positive excursions for a random walk with negative drift.

A central theme of the thesis is to extend results achieved for *iid* sequences to Markov chains. For this we need the corresponding theory of positive excursions for Markov additive processes (random walks controlled by a Markov chain) with negative drift. Along the way, new results for such processes are derived.

After giving a brief introduction in **Chapter 1** to biological sequences and a motivation for looking at excursions for Markov additive processes, the general theory of such processes and in particular the relevant excursion theory is treated in **Chapter 2**. Here we collect and customise all the results needed to treat the comparison of Markov chains in later chapters. Without much relation to the motivations from biological sequence analysis, we give in **Chapter 3** an extension of the excursion theory to incorporate heavy tails too. **Chapter 4** gives an introduction to the theory of alignment and structure of sequences. In particular, the focus in this chapter is on the construction of suitable models and test statistics for the problems mentioned above, e.g. testing if two sequences locally violate the hypothesis of being independent. We also discuss the formally similar problem of testing whether one sequence locally possesses 'more structure' than expected of a random sequence. In

Chapter 5 we derive the theoretical results about the local comparison of two independent Markov chains. These local comparisons result in a family of local scores. We derive a Poisson approximation of the number of scores exceeding a threshold, and based on this result we derive a Gumbel approximation of the maximal score. In **Chapter 6** we consider one Markov chain and compare it with itself corresponding to a specific local structural feature known as a stem-loop. We show that for this self-comparison, the number of scores exceeding a threshold can again be approximated by a Poisson distribution, and the maximal score can be approximated by a Gumbel distribution. Finally, in **Chapter 7** we give an example of how to use the results derived in Chapter 6 when searching large sequence databases for sequences with specific structures.

Resumé

I denne afhandling studerer vi sammenligning af sekvenser fra et endeligt alfabet og teorien for ekskursioner for Markov additive processer. Motivationen er primært anvendelser indenfor biologisk sekvensanalyse. Et hovedproblem er at teste om to sekvenser afviger fra hypotesen om at være *uafhængige* stokastiske sekvenser fra en specifik model, med en teststørrelse der afspejler den specifikke afvigelse fra modellen, som vi er interesseret i. Et tilsvarende problem, som introduceres i denne afhandling, er, hvorvidt en enkelt sekvens indeholder specifikke strukturelle egenskaber. For *lokale* afvigelser fra hypotesen, i det tilfælde hvor vi sammenligner uafhængige sekvenser af *iid* variable, er der opnået resultater af f.eks. Dembo et al. (1994*b*), som er baseret på, hvordan positive ekskursioner for random walks med negativ drift opfører sig.

Et af hovedtemaerne i afhandlingen er at udvide resultater opnået for *iid* sekvenser til Markovkæder. I den forbindelse får vi brug for teorien for positive ekskursioner for Markov additive processer (random walks kontrolleret af en Markovkæde) med negativ drift. Endvidere udvikles nye resultater for sådanne processer.

Efter en kort introduktion til biologisk sekvensanalyse i **kapitel 1** og en motivation for at betragte Markov additive processer, giver vi i **kapitel 2** en behandling af den generelle teori for sådanne processer – i særdeleshed den relevante teori for positive ekskursioner. I dette kapitel samler og tilpasser vi alle de resultater, der er behov for andetsteds for at sammenligne Markovkæder. Uden nogen synderlig relation til motivationen fra biologisk sekvensanalyse giver vi i **kapitel 3** en udvidelse af teorien til også at omfatte tunge haler. I **kapitel 4** gives en introduktion til teorien for alignments og strukturer for sekvenser. Vægten er i særdeleshed lagt på konstruktionen af passende modeller og teststørrelser for den slags problemer, som blev nævnt ovenfor. F.eks. testning af om to sekvenser lokalt afviger fra hypotesen om at være uafhængige. Vi diskuterer også det formelt tilsvarende problem, om en

sekvens lokalt har 'mere struktur' end man vil forvente for en tilfældig sekvens. I **kapitel 5** udledes de teoretiske resultater for lokal sammenligning af to uafhængige Markovkæder. Denne lokale sammenligning resulterer i en familie af lokale scoringer. Vi udleder en Poisson approksimation af antallet af scoringer, der overstiger en given grænse, og vi benytter dette resultat til at udlede en Gumbel approksimation af den maksimale score. I **kapitel 6** betragter vi istedet en enkelt Markovkæde og sammenligner den med sig selv, hvilket svarer til visse strukturelle egenskaber betegnet stem-loops. Vi viser, at for denne selv-sammenligning kan antallet af scoringer, der overstiger en given grænse, igen approksimeres med en Poisson fordeling, og den maksimale score kan approksimeres med en Gumbel fordeling. Endvidere giver vi i **kapitel 7** et eksempel på hvorledes resultaterne fra kapitel 6 kan anvendes i forbindelse med søgninger i store sekvensdatabaser efter sekvenser med specifikke strukturer.

Contents

Conventions and Notation

Notation is usually introduced when first needed. The following reference list may, however, be useful.

$\mathcal{D}(X)$	The distribution of a stochastic variable X.
$X \stackrel{\mathcal{D}}{=} Y$	Equality in distribution of X and Y.
$X_n \xrightarrow{\mathcal{D}} Y$	Convergence in distribution of X_n to Y.
$\overline{H}(x) = 1 - H(x)$	The right tail of a distribution function H.
$\|\|\nu\|\| = \sup_{f:\|f\|\leq 1} \int f d\nu$	The total variation norm of the signed measure ν.
$\|\|A\|\|$	The matrix of total variation norms when A is a matrix of signed measures.
v^t	The transposed of a vector v.
$\mathbb{1}$	The row vector $(1, \ldots, 1)^t$.
$1(B)$	The indicator function for the set B.
$\mathbb{E}(f(X); B) = \mathbb{E}(f(X)1(B))$	For a stochastic variable X and a set B.
$a(x) \sim b(x)$	$a(x)/b(x) \to 1$ for $x \to \infty$.
δ_x	The Dirac measure at x.
$\square$	End of proof.
$\diamond$	End of example.

Of course, effort has been put into choosing notation in accordance with common usage in the literature, and to choose notation consistent throughout the thesis. As this may lead to contradicting conditions some compromise between common usage and consistency is necessary. Most notably, Markov chains considered in the first part of the thesis are usually called $(J_n)_{n\geq 0}$, but in the last part, where we often consider two Markov chains, they are usually called $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$. Another important convention is that for a probability measure on $\mathbb{R}$ we do not distinguish in notation between the distribution function and the measure. Furthermore, we use

K, K_1, K_2 etc. to denote constants, whose value is not important and which may change throughout a proof.

Part I

Markov Controlled Excursions

1 Introduction

1.1 Comparison of biological sequences

DNA, RNA, and proteins are the fundamental building blocks for any living organism. The Central Dogma of biology states that DNA is the blueprint of the cell capable of copying itself, that DNA is *transcribed* into RNA, and that RNA is *translated* into functional proteins. These three ingredients, DNA, RNA, and protein, are all large molecules. DNA and RNA are made up of smaller parts called nucleotides or nucleic acids with DeoxyriboNucleic Acids making up DNA and the RiboNucleic Acids making up RNA. Proteins are made of amino acids. There are four different nucleic acids of each kind and 20 different amino acids. These molecules are usually connected in a linear sequence, and by representing each of them by a single letter, a DNA-, RNA-, or protein-molecule is very conveniently represented as a sequence of letters. Such sequences are known as *biological sequences*. The alphabets for representing the three different types of molecules are:

DNA	A, C, G, T
RNA	A, C, G, U
Protein	A, R, D, N, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V

But even though this provides a compact and useful way of representing the molecules, the letter sequences are only caricatures of what the molecules look like. The double stranded DNA is famous for its formation of the double helix and proteins are known to form complicated structures of a great variety. Moreover, these structures are of central importance for the way proteins work. Also RNA forms structures, and it was precisely the formation of certain RNA-structures that was the primary inspiration for the results developed in Chapter 6.

As for all good caricatures there is after all some truth contained in the sequence. The simple sequence representation of the molecules suggests a simple and highly useful way of comparing the molecules. One can basically try to match two sequences letter by letter and see how well they pair up. There are several reasons for being interested in making such comparisons, one being the desire to find evolutionary relation between species on a molecular level. Another is 'extrapolation of function by similarity', that is, if two sequences are very similar they are generally believed to have more or less the same function. Whatever the reason is, the methods used are very similar. Based on some assumptions about how to pair up letters in two sequences, one can find the so-called *optimal alignment* of the sequences very rapidly on a computer (Waterman 1995). These methods are known as *sequence alignment* and are by now classical methods in molecular biology. Perhaps the most important alignment tool developed is the BLAST search tool (Altschul et al. 1990, 1997) for *local alignment*. By local alignment we mean alignments that only take fractions of the sequences into account. This is extremely useful for searching databases containing lots of biological sequences in the hope of finding some sufficiently good matches.

The local alignment methods implemented in BLAST have also attracted a fair amount of theoretical interest. We mention Dembo et al. (1994*a*,*b*) and Siegmund & Yakir (2000, 2003) who have developed some of the important probabilistic underpinnings of BLAST. The purpose of probability theory is to give an estimate of the significance of the local alignments found from a sequence comparison. That is, to give an idea about whether the best local alignments found could just as well have occurred by chance when searching random sequences. BLAST usually does this by reporting a so-called E-value, which is the expected number of local alignments in random sequences with at least the same 'quality' as the best one found.

1.2 Random walks and excursions

The basic theoretical framework for discussing alignments is as follows. Given a set, or alphabet, E, two independent *iid* sequences $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ of random variables taking values in E and a so-called score function $f : E \times E \to \mathbb{R}$, we introduce the random walk $(S_n)_{n\geq 1}$ given by

$$S_n = \sum_{k=1}^{n} f(X_k, Y_k).$$

The variable S_n can be thought of as a comparison of the first n variables from the X-sequence with the first n variables from the Y-sequence using the function f. The increment $S_m - S_k$ of the random walk from $k+1$ to m corresponds to comparing only

the variables from $k+1$ to m, and the maximum of these increments for $k \leq m \leq n$ is the largest *local* comparison score up to n. It is also the largest positive increment of the random walk up to n, and we will refer to this as the *maximal positive excursion* of the random walk up to n. This is in itself an interesting stochastic variable[1], but what makes it even more complicated is that we also want to shift the sequences alongside each other. That is, in addition to $(S_n)_{n\geq 1}$ we also want to introduce the random walks $(S_n^T)_{n\geq 1}$ defined by

$$S_n^T = \sum_{k=1}^{n} f(X_k, Y_{k+T})$$

for $T \geq 1$, their corresponding increments and maximal positive excursions. All the random walks have the same distribution but they are dependent, which makes the analysis of e.g. the total number of positive excursions exceeding a threshold harder.

Although shifting makes the problem more difficult, it is first of all important to understand the behaviour of the random walk. Interestingly, this was originally done with a queueing application in mind (Iglehart 1972). Later, the behaviour of the maximal positive excursion was reconsidered by Karlin & Dembo (1992) in a molecular comparison context and extended to also deal with Markov controlled random walks.

1.3 Markov additive processes

One of the new results presented in this thesis is an extension of the theory of local alignments to deal with aligning independent Markov chains instead. If $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ are Markov chains instead of *iid* sequences, the process $(S_n)_{n\geq 1}$ is known as a Markov controlled random walk or a Markov additive process. Such processes can be dealt with in much the same way as random walks, and one can, in particular, derive similar results as for random walks about the maximal positive excursions as done by Karlin & Dembo (1992).

In a Markov setup it may also be reasonable to choose a score function f that not only compares the sequences letter by letter, but instead compares the sequences transition by transition. That is, the score function f is instead defined on $E^2 \times E^2$ and

$$S_n = \sum_{k=1}^{n-1} f((X_n, X_{n+1}), (Y_n, Y_{n+1})).$$

[1]to make it really interesting the random walk must have negative drift

Indeed, as we derive in Chapter 4, there is in the Markov setup a natural score function of this form. Moreover, even if we just want to generalise our score function to transition scoring as above but want to keep the *iid* assumption, we are forced into the realm of Markov additive processes anyway. In our treatment of local alignments of Markov chains in Chapter 5 we restrict ourself to score functions that compare letter with letter, but, as discussed in Chapter 5, there is no loss of generality in doing so since we can always stack the Markov chains. The restriction is done for notational convenience only. Nevertheless, in the treatment of the positive excursions for the Markov additive processes presented in Chapter 2 we treat scoring of transitions in generality. The reason is for instance that certain representations of constants becomes substantially more complicated if we are forced to stack the Markov chain before applying the representations. Moreover, without much extra effort we can and will treat Markov additive processes in general and not 'just' sums of functions of a Markov chain, cf. the definition of a Markov additive process in Chapter 2.

1.4 Integer scoring

The score function f used is often derived by the methods presented in Chapter 4, for which the function naturally takes values on the real line. The function does, however, always take a finite number of values, which by the limited precision necessarily are located on a lattice. For the theoretical results it is necessary to take this lattice phenomenon into account, though it is quite often ignored and most likely matters very little. To make the lattice phenomenon obvious we choose to assume that f takes values in $\mathbb{Z}$ instead. Moreover, computations with integers are often faster on computers than floating point computations, so it is also tradition in real applications to construct integer versions of the score function that one wants to use. For a given real valued f, simply rounding f to the nearest integer may produce a very dull function, but if we multiply f by some constant $\kappa \geq 1$ and round to the nearest integer we can obtain something more interesting. Define the κ-integer version of f as

$$I_\kappa(f) := \lfloor \kappa f + 0.5 \rfloor,$$

that is, $I_\kappa(f)$ is the rounding of κf to the nearest integer (rounding 0.5 up). Taking $\kappa = 10^k$ yields a score function equivalent to f with k digits. It is also obvious that a multiplication of f just corresponds to scaling the random walk/Markov additive process. Thus up the scaling factor (and rounding effect) we will obtain the same results whether we use f or $I_\kappa(f)$. The larger κ is, the higher the precision will be, the larger the absolute values of f will be and the less will the lattice effect matter. Having f take too large absolute values may lead to computational problems for

some of the algorithms discussed in Chapter 2, which are actually based on the lattice effect. Therefore we use a very moderate κ in Chapter 7 in a concrete application.

2

Excursion Theory for Markov Additive Processes

2.1 Markov additive processes

A Markov additive process is a random walk, whose increments are controlled by a Markov chain. Thus assume that $(J_n)_{n\geq 0}$ is a Markov chain on a finite state space E with transition probability matrix $P = (P_{ij})_{i,j\in E}$, and that $(Z_n)_{n\geq 1}$ conditionally on $(J_n)_{n\geq 0}$ is a sequence of independent stochastic variables with values in $\mathbb{R}$. Furthermore, with $(H_{ij})_{i,j\in E}$ a matrix of probability measures assume that the conditional distribution of Z_n given $(J_n)_{n\geq 0}$ is $H_{J_{n-1}J_n}$. We call $(H_{ij})_{i,j\in E}$ the increment distributions. Defining $S_0 = 0$ and $S_n = \sum_{k=1}^{n} Z_k$ for $n \geq 1$, the bivariate process $(J_n, S_n)_{n\geq 0}$ (and sometimes just $(S_n)_{n\geq 0}$) is called a Markov additive process – or a MAP for short. We will use F to denote the matrix of positive measures given by $F_{ij} = H_{ij}P_{ij}$, in which case the transition probabilities for the bivariate Markov process $(J_n, Z_n)_{n\geq 1}$ are given by

$$\mathbb{P}(Z_n \in A, J_n = j \mid Z_{n-1}, J_{n-1} = i) = F_{ij}(A),$$

for $A \in \mathbb{B}$ a Borel set. Throughout we use $\mathbb{P}_\nu$ to denote a probability measure under which the Markov chain has initial distribution ν on E, and in particular when $\nu = \delta_i$ we write $\mathbb{P}_i$ instead of $\mathbb{P}_{\delta_i}$.

In the statistical literature, the process $(Z_n)_{n\geq 1}$ is known as a hidden Markov chain, and the major concern is estimation of parameters given that one observes the Z-process only. This is, however, not an important issue in this thesis. Rather we are interested in the behaviour of the process $(S_n)_{n\geq 0}$, and especially we focus on

the maximal positive excursions, or maximal positive increments, the process makes under the assumption that it has an overall negative drift.

This chapter introduces a variety of concepts, tools and results, which are all more or less well developed in the literature. This will serve a three fold purpose. First of all the reader will have a reference at hand for later chapters, where we will use the tools and results. Second, in Chapter 3 new results comparable with some of those presented in this chapter will be derived, and it may be useful to be able to compare results and techniques of proof. Third, important constants appearing have some quite complicated and not entirely intuitive representations, so by giving complete proofs we can show how the constants appear and give useful – and hopefully insightful – representations. Though the author does not claim any originality for the main part of results presented in this chapter, it has been decided not to clutter up the text with references. Instead relevant references can be found in the notes at the end of the chapter. References are, however, given in the text for those results that are stated without proof.

Before continuing the general treatment of MAPs, we consider a special case – the Markov renewal process – which becomes useful several times in the following.

Example 2.1.1 If $H_{ij}((0,\infty)) = 1$ for all $i, j \in E$, the process $(S_n)_{n\geq 0}$ is known as a Markov renewal process. As for ordinary renewal processes, a Blackwell-type renewal theorem exists for Markov renewal processes. Assume that the matrix P of transition probabilities for the underlying Markov chain is irreducible. The renewal process is then either concentrated on a minimal lattice or not, cf. Çinlar (1975), Section 10.2. Let the invariant probability vector for P be ν, then with

$$U_{ij}(A) = \sum_{n=0}^{\infty} \mathbb{P}_i(J_n = j, S_n \in A)$$

the Markov renewal kernel and $\mu = \sum_{ij} \int u F_{ij}(\mathrm{d}u)\nu_i$, it holds for $t \to \infty$ that

$$\begin{aligned} U_{ij}([t, t+s)) &\to s\frac{\nu_j}{\mu} \quad \text{in the non-lattice case,} \\ U_{ij}(\{t\}) &\to \frac{\nu_j}{\mu} \quad \text{in the lattice case with } t \text{ on the lattice.} \end{aligned}$$

Renewal equations can also be generalised to the Markov renewal setup. Assume that $z_i : [0,\infty) \to \mathbb{R}$ for $i \in E$ are a given set of functions. A vector $(Z_i)_{i\in E}$ of functions $Z_i : [0,\infty) \to \mathbb{R}$ fulfills a Markov renewal equation (or a system of renewal equations) if

$$Z_i(t) = z_i(t) + \sum_{k\in E} \int_0^t Z_k(t-u) F_{ik}(\mathrm{d}u), \ i \in E.$$

If the z_i's are all directly Riemann integrable, the Markov key renewal theorem holds:

$$Z_i(t) \to \frac{1}{\mu} \sum_{k \in E} \nu_k \int_0^\infty z_k(u)(\mathrm{d}u) \tag{2.1}$$

for $t \to \infty$ – along the lattice if necessary. Both of these results can be found in Asmussen (2003), Section VII.4. Chapter 10 in Çinlar (1975) is a more classical reference to this subject. ◇

We will assume throughout the rest of this chapter that:

1. The Markov chain is irreducible and aperiodic with invariant distribution π.
2. All the measures H_{ij} have first moments and defining
$$\mu_{ij} := \int x H_{ij}(\mathrm{d}x)$$
we assume that
$$\mu := \sum_{i,j} \pi_i P_{ij} \mu_{ij} = \sum_{i,j} \pi_i \int x F_{ij}(\mathrm{d}x) < 0.$$
3. There exists a finite sequence $j_0, \ldots, j_{n-1}$ satisfying $P_{j_{k-1},j_k} > 0$ for $k = 1, \ldots, n-1$ and $P_{j_{n-1},j_0} > 0$ such that
$$\mathbb{P}_{j_0}\left(\sum_{k=1}^n Z_k > 0 \mid J_1 = j_1, \ldots, J_{n-1} = j_{n-1}, J_n = j_0\right) > 0.$$

Condition 2 above is referred to as *the negative drift assumption* and Condition 3 as the *non-degeneracy condition*. Usually a finite sequence $j_0, \ldots, j_{n-1}$ satisfying $P_{j_{k-1},j_k} > 0$ for $k = 1, \ldots, n-1$ and $P_{j_{n-1},j_0} > 0$ is called a cycle w.r.t. the transition probability matrix P. Thus Condition 3 says that there should exist a cycle w.r.t. P along which there is positive probability of observing a strictly positive increment for the MAP. This corresponds to assuming for an ordinary random walk that there is positive probability for observing a strictly positive increment.

It is easily verified that the process $(J_n, Z_n)_{n \geq 1}$ is ergodic, so by the negative drift assumption and the Ergodic Theorem $S_n \to -\infty$ a.s. for $n \to \infty$.

The time reversed MAP, which will be useful too, is the MAP where the Markov chain has transition probabilities $\overleftarrow{P}$,

$$\overleftarrow{P}_{ij} = \frac{\pi_j P_{ji}}{\pi_i},$$

and the increments have conditional distributions $\overleftarrow{H}_{ij} = H_{ji}$. It is thus given by simply time reversing the Markov chain and reversing the increments. We use $(\overleftarrow{J}_n, \overleftarrow{S}_n)_{n \geq 0}$ to denote a time-reversed MAP.

2.2 Ladder variables and Wiener-Hopf factorisation

Define the stopping times

$$\begin{aligned}\tau_- &= \inf\{n > 0 \mid S_n \leq 0\} \quad \text{and} \\ \tau_+ &= \inf\{n > 0 \mid S_n > 0\},\end{aligned}$$

which are known respectively as the first descending ladder time and the first (strict) ascending ladder time. The distribution of (J_{τ_-}, S_{τ_-}) and the distribution of (J_{τ_+}, S_{τ_+}) (on $(\tau_- < \infty)$ and $(\tau_+ < \infty)$ respectively) are given by two matrices G_- and G_+ of positive measures;

$$\begin{aligned}G_{-,ij}(A) &= \mathbb{P}_i(J_{\tau_-} = j, S_{\tau_-} \in A, \tau_- < \infty) \quad \text{and} \\ G_{+,ij}(A) &= \mathbb{P}_i(J_{\tau_+} = j, S_{\tau_+} \in A, \tau_+ < \infty)\end{aligned}$$

with $A \in \mathbb{B}$ a Borel subset of $\mathbb{R}$. The matrix G_- is the matrix of descending ladder height distributions and G_+ is the matrix of ascending ladder height distributions. Under our general assumptions, $\tau_- < \infty$ a.s. due to the negative drift condition but $\mathbb{P}_i(\tau_+ = \infty) > 0$ may occur.

Often the sequences $(\tau_-(n))_{n\geq 0}$ and $(\tau_+(n))_{n\geq 0}$ of stopping times defined by

$$\begin{aligned}\tau_-(n) &= \inf\{k > \tau_-(n-1) \mid S_k \leq S_{\tau_-(n-1)}\}, \quad \tau_-(0) = 0 \\ \tau_+(n) &= \inf\{k > \tau_+(n-1) \mid S_k > S_{\tau_+(n-1)}\}, \quad \tau_+(0) = 0\end{aligned}$$

are useful. We have that $\tau_-(1) = \tau_-$ and $\tau_+(1) = \tau_+$ and the sequences are known as the descending and ascending ladder epochs. Due to the negative drift, $\tau_-(n) < \infty$ a.s. for all n and $\tau_+(n) = \infty$ eventually a.s.

The matrices G_- and G_+ are elements in a convolution algebra of matrices of signed measures on $\mathbb{R}$. For two matrices A and B of signed measures, the convolution product of A and B is simply

$$(A * B)_{ij} = \sum_k A_{ik} * B_{kj}.$$

Note that the convolution product of matrix measures is non-commutative. Note also that the matrix $I = \text{diag}\{\delta_0, \ldots, \delta_0\}$, being the diagonal matrix with Dirac measures at zero in the diagonal, serves as the identity under convolution. For any matrix A of signed measures we use the notation $||A||$ to denote the matrix of total variation norms of the measures in A. For a matrix, A, of positive measures, $||A||$ is then the matrix of total masses. With this notation $||G_-||$ is recognised as the matrix of transition probabilities for the Markov chain $(J_{\tau_-(n)})_{n\geq 0}$.

For the time reversed MAP we can likewise define the ladder times and matrices of ladder height distributions, which we naturally denote $\overleftarrow{G}_-$ and $\overleftarrow{G}_+$.

Theorem 2.2.1 *With* ${}^{\#}G_+ = ({}^{\#}G_{+,ij})_{i,j\in E}$ *defined by*

$$ {}^{\#}G_{+,ij}(A) = \frac{\pi_j \overleftarrow{G}_{+,ji}(A)}{\pi_i} $$

and ${}^{\#}G_-$ *defined similarly, the Wiener-Hopf factorisation identity*

$$ I - F = (I - {}^{\#}G_+) * (I - G_-) = (I - {}^{\#}G_-) * (I - G_+) \tag{2.2} $$

holds.

The proof is in fact quite easy but is skipped, cf. Theorem XI.2.12 in Asmussen (2003). By evaluating these matrix measures on $\mathbb{R}$, the Wiener-Hopf identity amounts to the ordinary matrix equation

$$ I - P = I - ||F|| = (I - ||{}^{\#}G_+||)(I - ||G_-||) = (I - ||{}^{\#}G_-||)(I - ||G_+||) \tag{2.3} $$

with $I = \mathrm{diag}(1, \ldots, 1)$ also denoting the ordinary identity matrix.

The Wiener-Hopf identity has many useful consequences as the following sections as well as Chapter 3 will show. As an immediate but nevertheless non-trivial observation, it follows from (2.3) that the stochastic matrix $||G_-||$ has only one communication class – and maybe some transient states. Indeed if it was reducible, i.e. had two disjoint communication classes, it would have a right eigenvector with eigenvalue 1 differing from $\mathbb{1} = (1, \ldots, 1)^t$ and hence so would P contradicting the irreducibility of P. Thus the Markov chain $(J_{\tau_-(n)})_{n\geq 0}$ may have some transient states, but there is only one irreducibility class. In particular, there exists a state i_0 to which the Markov chain returns infinitely often independent of the initial distribution. The Markov chain is thus *Harris recurrent*.

2.3 Exponential change of measure

Define the matrix $\Phi(\theta)$ for $\theta \in \mathbb{R}$ by

$$ \Phi(\theta)_{ij} = \int \exp(\theta x) F_{ij}(\mathrm{d}x) = P_{ij} \int \exp(\theta x) H_{ij}(\mathrm{d}x) = \mathbb{E}_i(\exp(\theta Z_1); J_1 = j). $$

We will assume for the rest of this chapter that $\Phi(\theta)_{ij} < \infty$ for all $i, j \in E$ and all $\theta \in \mathbb{R}$. It may seem unnecessarily restrictive to assume $\Phi(\theta)_{ij} < \infty$ for all $\theta \in \mathbb{R}$

instead of just for some $\theta > 0$. The main application in mind is the case with H_{ij} all having compact – indeed finite – support, in which case $\Phi(\theta)_{ij}$ is clearly finite for all $\theta \in \mathbb{R}$. All results derived in this chapter hold if just $\Phi(\theta)_{ij} < \infty$ for some $\theta > 0$ with a few obvious modifications.

The matrix $\Phi(\theta)$ is a matrix with positive entries, which is irreducible due to the irreducibility of P. Perron-Frobenius theory tells us that the spectral radius of $\Phi(\theta)$ is a simple eigenvalue with unique (up to scaling) left and right eigenvectors. For a matrix A we denote the spectral radius of A by $\mathrm{spr}(A)$. Define the function $\varphi : \mathbb{R} \to \mathbb{R}$ by

$$\varphi(\theta) = \mathrm{spr}(\Phi(\theta)). \tag{2.4}$$

The function φ serves in this setup the same purpose as the Laplace transform of the increments of a random walk usually does, and we put $\psi(\theta) = \log \varphi(\theta)$ corresponding to the cumulant generating function. It is a strictly convex C^∞ function.

The convexity of functions like ψ (and hence the log-convexity of φ) goes back to Kingman (1961). We will need strict convexity in certain situations, for which we state the following criterion as a consequence of Theorem 4 in O'Cinneide (2000). If

$$\theta \mapsto \log \Phi(\theta)_{ij} \text{ is strictly convex for some } i, j \in E \tag{2.5}$$

then ψ is strictly convex. If (2.5) is *not* fulfilled, all the distributions H_{ij} are degenerate, with, say, $H_{ij} = \delta_{f(i,j)}$ for some function $f : E \times E \to \mathbb{R}$. In this case ψ is strictly convex if there exists a cycle $i_0, \ldots, i_{n-1}$ w.r.t. P such that

$$\sum_{k=1}^{n-1} f(i_{k-1}, i_k) \neq \mu. \tag{2.6}$$

This holds whether or not Condition 2 and 3, as otherwise assumed, are fulfilled. But if Condition 2 and 3 are fulfilled we observe that ψ is strictly convex.

Differentiability is a consequence of the Implicit Function Theorem. To be precise, let $l^\theta = (l_i^\theta)_{i \in E}$ be the left (row) eigenvector for $\Phi(\theta)$ corresponding to the eigenvalue $\varphi(\theta)$ – normalised so that it sums to 1. Then $(l^\theta, \varphi(\theta))$ is by Perron-Frobenius theory the unique solution to the equations

$$\begin{aligned} l^\theta \Phi(\theta) &= \varphi(\theta) l^\theta, \\ l^\theta \mathbb{1} &= 1 \end{aligned}$$

in the quadrant where $\varphi > 0$ and $l_i > 0$. Likewise there is an up to scaling unique, right (column) eigenvector, $r^\theta = (r_i^\theta)_{i \in E}$, with strictly positive entries, which we will choose to normalise by $l^\theta r^\theta = 1$. Introducing the map

$$\begin{aligned} g_\theta : (0, \infty)^{|E|+1} &\to \mathbb{R}^{|E|+1} \\ (l, \varphi) &\mapsto (l(\Phi(\theta) - \varphi I), l\mathbb{1} - 1), \end{aligned}$$

$(l^\theta, \varphi(\theta))$ is (the unique) solution to $g_\theta(l, \varphi) = 0$. The derivative of g_θ is

$$Dg_\theta(l, \varphi) = \left\{ \begin{array}{cc} \Phi(\theta) - \varphi I & \mathbb{1} \\ -l & 0 \end{array} \right\}.$$

Assume that the (column) vector $(v^t, v_0)^t$ (with v a column vector of length $|E|$) is in the kernel of $Dg_\theta(l^\theta, \varphi(\theta))$, then

$$\begin{aligned} \Phi(\theta)v &= \varphi(\theta)v - v_0\mathbb{1}, \\ l^\theta v &= 0. \end{aligned}$$

We observe that

$$0 = \varphi(\theta)l^\theta v = l^\theta \Phi(\theta)v = \varphi(\theta)l^\theta v - v_0 l^\theta \mathbb{1} = -v_0.$$

But when $v_0 = 0$ the vector v must be 0 or a right Perron-Frobenius eigenvector with eigenvalue $\varphi(\theta)$, which must hence have strictly positive entries due to irreducibility. The fact that $l^\theta v = 0$ contradicts the last possibility and $v = 0$. By uniqueness of the solution $(l^\theta, \varphi(\theta))$ in the open quadrant, the Implicit Function Theorem implies that $\varphi(\theta)$ as well as l^θ (and hence also r^θ) are C^∞ in θ.

Differentiating the equation

$$\varphi(\theta) r_i^\theta = \sum_j \Phi(\theta)_{ij} r_j^\theta$$

gives for $\theta = 0$ that

$$\partial_\theta \varphi(0) r_i^0 + \varphi(0)\partial_\theta r_i^0 = \sum_j \partial_\theta \Phi(0)_{ij} r_j^0 + P_{ij}\partial_\theta r_j^0.$$

Since obviously $\varphi(0) = 1$, $l^0 = \pi$ and $r^0 = \mathbb{1} = (1, \ldots, 1)^t$, multiplication by π_i and summation yield that

$$\partial_\theta \varphi(0) \ (= \partial_\theta \psi(0)) = \sum_{ij} \pi_i P_{ij} \mu_{ij} = \mu. \tag{2.7}$$

This is one reason that φ serves as a natural generalisation of the Laplace transform.

We define a process $(L_n^\theta)_{n \geq 0}$ for $\theta \geq 0$ by

$$L_n^\theta = \frac{r_{J_n}^\theta}{\varphi(\theta)^n \, r_{J_0}^\theta} \exp(\theta S_n) = \frac{r_{J_n}^\theta}{r_{J_0}^\theta} \exp(\theta S_n - n\psi(\theta)).$$

With $(\mathcal{F}_n)_{n\geq 0}$ the filtration of σ-algebras generated by the MAP $(J_n, S_n)_{n\geq 0}$, we find that

$$\begin{aligned}
\mathbb{E}_i(L_n^\theta \mid \mathcal{F}_{n-1}) &= \mathbb{E}_i\left(\frac{r^\theta_{J_n}}{\varphi(\theta)^n\, r^\theta_{J_0}}\exp(\theta S_n)\;\middle|\;(J_{n-1}, S_{n-1})\right)\\
&= \frac{\exp(\theta^* S_{n-1})}{\varphi(\theta)^n r_i^\theta}\mathbb{E}_i\left(r^\theta_{J_n}\exp(\theta Z_n)\;\middle|\;J_{n-1}\right)\\
&= \frac{L^\theta_{n-1}}{\varphi(\theta)\, r_{J_{n-1}}}\sum_j \mathbb{E}_{J_{n-1}}(\exp(\theta Z_1); J_1 = j)P_{J_{n-1}j}r_j^\theta\\
&= L^\theta_{n-1}\frac{(\Phi(\theta)r)_{J_{n-1}}}{\varphi(\theta)\, r_{J_{n-1}}} = L^\theta_{n-1}.
\end{aligned}$$

Together with $\mathbb{E}_i(L_0^\theta) = 1$ and $L_n^\theta > 0$ this shows that $(L_n^\theta, \mathcal{F}_n)_{n\geq 0}$ is a positive martingale, which thus are the densities for a sequence of probability measures on the filtration $(\mathcal{F}_n)_{n\geq 0}$ of σ-algebras. Such a process is sometimes referred to as a likelihood process. If the processes considered are defined on a suitable space, e.g. the canonical one $E^{\mathbb{N}_0}\times\mathbb{R}^{\mathbb{N}_0}$, there exists a unique measure $\mathbb{P}_i^\theta$ such that

$$\frac{\mathrm{d}\mathbb{P}_i^\theta|_{\mathcal{F}_n}}{\mathrm{d}\mathbb{P}_i|_{\mathcal{F}_n}} = L_n^\theta.$$

The measure $\mathbb{P}_i^\theta$ is called the exponentially changed or tilted measure.

Since the eigenvector fractions are bounded above and bounded away from 0, the identity $\mathbb{E}_i(L_n^\theta) = 1$ valid for all θ implies that

$$\frac{1}{n}\log\mathbb{E}_i\left(\exp(\theta S_n)\right) \to \psi(\theta) \tag{2.8}$$

for $n\to\infty$. Thus $n\psi$ can be viewed as an asymptotic cumulant generating function for S_n.

The tilted measure defined in the following lemma will be of particular importance.

Lemma 2.3.1 *There exists a unique $\theta^* > 0$ solving the equation $\varphi(\theta) = 1$ ($\psi(\theta) = 0$). For this θ^* it holds that $\partial_\theta\varphi(\theta^*) > 0$.*

Proof: Since $\varphi(0) = 1$ and $\partial_\theta\varphi(0) = \mu < 0$ there exists a $\theta^* > 0$ with $\varphi(\theta^*) = 1$ if $\varphi(\theta)\to\infty$ for $\theta\to\infty$. Due to convexity θ^* is then a unique solution and necessarily $\partial_\theta\varphi(\theta^*) > 0$ holds. Hence we just have to show that $\varphi(\theta)\to\infty$. But due to the non-degeneracy condition there exists a sequence $j, j_1, \ldots, j_{n-1}$ such that

$$\Phi(\theta)^n_{jj} \geq \mathbb{E}_j\left(\exp\left(\theta\sum_{k=1}^n Z_k\right); \sum_{k=1}^n Z_k > 0, J_1 = j_1, \ldots, J_{n-1} = j_{n-1}, J_n = j\right) \to \infty$$

for $\theta \to \infty$. In particular, defining $A(\theta)$ as the matrix with 0 everywhere except for $A(\theta)_{jj} = \Phi(\theta)^n_{jj}$, we have that $\Phi(\theta)^n \geq A(\theta)$ coordinatewise and

$$\varphi(\theta) = \mathrm{spr}(\Phi(\theta)) = \mathrm{spr}(\Phi(\theta)^n)^{1/n} \geq \mathrm{spr}(A(\theta))^{1/n} = \left(\Phi(\theta)^n_{jj}\right)^{1/n} \to \infty.$$

□

For this particular θ^* we let $P^* = P^{\theta^*}$, $r^* = r^{\theta^*}$, etc. and we will denote the exponential changed measure by $\mathbb{P}^*_i$. Expectations w.r.t. to $\mathbb{P}^*_i$ will be denoted $\mathbb{E}^*_i$.

It is straight forward to verify that under $\mathbb{P}^\theta_i$ the process $(J_n, S_n)_{n\geq 0}$ is still a MAP (the tilted MAP) with transition probabilities P^θ and increment distributions H^θ given by

$$\begin{aligned} P^\theta_{ij} &= \frac{r^\theta_j}{\varphi(\theta)\, r^\theta_i}\Phi(\theta)_{ij} \\ \frac{\mathrm{d}H^\theta_{ij}}{\mathrm{d}H_{ij}}(x) &= \frac{P_{ij}}{\Phi(\theta)_{ij}}\exp(\theta x). \end{aligned}$$

For $\theta_0 \in \mathbb{R}$ we can likewise define $\Phi^{\theta_0}(\theta)$ for the tilted MAP as

$$\Phi^{\theta_0}(\theta)_{ij} = P^{\theta_0}_{ij}\int \exp(\theta x) H^{\theta_0}_{ij}(\mathrm{d}x)$$

and $\varphi^{\theta_0}(\theta) = \mathrm{spr}(\Phi^{\theta_0}(\theta))$. Simple computations reveal that

$$\varphi^{\theta_0}(\theta) = \frac{\varphi(\theta+\theta_0)}{\varphi(\theta_0)} \quad \text{or} \quad \psi^{\theta_0}(\theta) = \psi(\theta+\theta_0) - \psi(\theta_0),$$

and from (2.7) it follows that the tilted MAP has drift $\partial_\theta\psi(\theta_0)$. The tilted MAP corresponding to θ^* thus have positive drift $\partial_\theta\varphi(\theta^*) = \partial_\theta\psi(\theta^*)$.

A highly fruitful observation is that for any stopping time τ, L^θ_τ is still a likelihood, i.e. we have that on $(\tau < \infty)$

$$\frac{\mathrm{d}\mathbb{P}^\theta_i|_{\mathcal{F}_\tau}}{\mathrm{d}\mathbb{P}_i|_{\mathcal{F}_\tau}} = L^\theta_\tau. \tag{2.9}$$

As a consequence, for $\mathbb{E}_i(\tau) < \infty$, differentiating the equation $\mathbb{E}_i(L^\theta_\tau) = 1$, putting $\theta = 0$ and doing some simple algebra yield the following generalisation of Wald's identity.

Theorem 2.3.2 *If* $\mathbb{E}_i(\tau) < \infty$ *then*

$$\mathbb{E}_i(S_\tau) = \mu\mathbb{E}_i(\tau) + \mathbb{E}_i\partial_\theta r^0_{J_\tau} - \partial_\theta r^0_i$$

where $\mu = \sum_{ij}\pi_i P_{ij}\mu_{ij} = \partial_\theta\varphi(0)$.

Remark 2.3.3 *Observe that this generalisation of Wald's identity contains the slightly mysterious term* $\mathbb{E}_i \partial_\theta r^0_{J_\tau} - \partial_\theta r^0_i$. *For a given initial distribution* ν *we get "classical Wald"*

$$\mathbb{E}_\nu(S_\tau) = \mu \mathbb{E}_\nu(\tau) \quad \text{if and only if} \quad \mathbb{E}_\nu \partial_\theta r^0_{J_0} = \mathbb{E}_\nu \partial_\theta r^0_{J_\tau},$$

which holds if e.g. $J_0 \stackrel{\mathcal{D}}{=} J_\tau$ *under* $\mathbb{P}_\nu$.

2.4 Reflection, excursions and extremes

For the additive process $S_0 = 0,\ S_n = \sum_{k=1}^n Z_k$, the reflection at the zero barrier is defined as the process $(T_n)_{n\geq 0}$ given recursively by

$$T_n = (T_{n-1} + Z_n)^+, \quad T_0 = 0,$$

where $x^+ = \max(0, x)$. This process can also be represented as

$$T_n = S_n - \min_{0\leq k\leq n} S_k, \tag{2.10}$$

which is seen by observing that the r.h.s. above fulfills the recursion. A useful consequence of the last representation of $(T_n)_{n\geq 0}$ is that

$$\max_{0\leq k\leq m\leq n} S_m - S_k = \max_{0\leq m\leq n} T_m. \tag{2.11}$$

Thus the maximum of all partial sums $\sum_{r=k+1}^m X_r$ for $k \leq m \leq n$ equals the maximum of T_m for $m \leq n$. In Chapter 5 and Chapter 6 we take an interest in the maximum over such partial sums, in which case (2.11) gives both computational as well as analytic advantages.

The process $(T_n)_{n\geq 0}$ can be decomposed into regenerative cycles. It starts at zero, and at each descending ladder epoch for the additive process it returns to zero again. If we fix a state $i_0 \in E$ such that $\widetilde{\sigma} = \inf\{n > 0 \mid J_{\tau_-(n)} = i_0\}$ is finite due to Harris recurrence of $(J_{\tau_-(n)})_{n\geq 0}$, then

$$\sigma = \inf\{n \geq 0 \mid T_n = 0, J_n = i_0\} = \sum_{n=1}^{\widetilde{\sigma}} \tau_-(n), \tag{2.12}$$

is finite and a regeneration time for the Markov process $(J_n, T_n)_{n\geq 0}$. We define the *cycle maximum* as

$$\mathcal{M}_\sigma = \max_{1\leq n\leq \sigma} T_n$$

and want to determine the asymptotic behaviour of $\mathbb{P}_{i_0}(\mathcal{M}_\sigma > u)$ for $u \to \infty$. Having established the asymptotic behaviour of the cycle maximum it is rather easy to derive the asymptotic behaviour of the running maximum

$$\mathcal{M}_n = \max_{0 \le k \le n} T_k,$$

since this is essentially a maximum over independent cycles.

To determine how $\mathbb{P}_{i_0}(\mathcal{M}_\sigma > u)$ behaves we consider first

$$\mathcal{M}_{\tau_-} = \max_{1 \le n \le \tau_-} S_n$$

and determine the behaviour of $\mathbb{P}_i(\mathcal{M}_{\tau_-} > u)$ for $u \to \infty$.

Define $\tau(u) = \inf\{n \ge 0 \mid S_n > u\}$ and use the exponential change of measure to get that

$$\begin{aligned} \mathbb{P}_i(\mathcal{M}_{\tau_-} > u) &= \mathbb{P}_i(\tau_- > \tau(u)) \\ &= \mathbb{E}_i^*((L^*_{\tau(u)})^{-1}; \tau_- > \tau(u)) \\ &= r_i^* \exp(-\theta^* u) \mathbb{E}_i^* \left(\frac{1}{r^*_{J_{\tau(u)}}} \exp\left(-\theta^*(S_{\tau(u)} - u)\right); \tau_- > \tau(u) \right). \end{aligned} \tag{2.13}$$

With $K = \max_i r_i^* / \min_j r_j^*$, the Lundberg-type inequality

$$\mathbb{P}_i(\mathcal{M}_{\tau_-} > u) \le K \exp(-\theta^* u) \tag{2.14}$$

follows, giving that the probability as a function of u must decay at least exponentially fast. To establish the correct asymptotic behaviour we need to control the expectation on the right hand side in (2.13) for $u \to \infty$. It will be established as a central result below that under $\mathbb{P}_i^*$ the pair of stochastic variables $(J_{\tau(u)}, S_{\tau(u)} - u)$ converges in distribution to a limit, which can be given explicitly by the ladder height distributions, and which is asymptotically independent of the events $(\tau_- > \tau(u))$.

For the purpose of this thesis we will only need to consider the case where H_{ij} are concentrated on a lattice[1]. We will assume throughout that this (minimal) lattice is $\mathbb{Z}$. Any limits will therefore take place in $\mathbb{Z}$.

Theorem 2.4.1 *Under $\mathbb{P}_i^*$ we have for $u \to \infty$ that*

$$(J_{\tau(u)}, S_{\tau(u)-u}) \xrightarrow{\mathcal{D}} (J_\infty, B_\infty).$$

[1] in contrast to the usual assumption, which is that H_{ij} is *not* concentrated on a lattice.

for some bivariate stochastic variable (J_∞, B_∞). *Furthermore,*

$$\mathbb{P}_i(\mathcal{M}_{\tau_-} > u) \sim c\, c_i \exp(-\theta^* u)$$

with $c_i = r_i^* \mathbb{P}_i^*(\tau_- = \infty)$ *and*

$$c := \mathbb{E}\left(\frac{1}{r^*_{J_\infty}} \exp(-\theta^* B_\infty)\right).$$

Proof: Under $\mathbb{P}_i^*$ the process $(J_n, S_n)_{n\geq 0}$ is a MAP with average drift $\partial_\theta \psi(\theta^*) > 0$ so $S_n \to \infty$ a.s. by the ergodic theorem. Hence $\tau_+, \tau(u) < \infty$ almost surely. The key observation is that for fixed $z \in \mathbb{N}$, the probabilities

$$A_{ij}(u) := \mathbb{P}_i^*(S_{\tau(u)} - u = z, J_{\tau(u)} = j),\ u \in \mathbb{N}_0,$$

fulfill for each $j \in E$ the Markov renewal equation

$$A_{ij}(u) = a_{ij}(u) + \sum_k \sum_{v=1}^{u} A_{kj}(u-v)\mathbb{P}_i^*(S_{\tau+} = v, J_{\tau_+} = k),\ i \in E,$$

where $a_{ij}(u) = \mathbb{P}_i^*(S_{\tau_+} = z + u, J_{\tau_+} = j)$. This is easily seen by dividing according to whether $\tau_+ = \tau(u)$ or $\tau_+ < \tau(u)$ and in the last case conditioning on the value of (J_{τ_+}, S_{τ_+}).

It is seen by the definition that $\left(\mathbb{P}_i^*(S_{\tau+} \in \cdot, J_{\tau_+} = j)\right)_{i,j\in E}$ is the matrix of ladder height distributions *under the exponentially changed measure*, which we will naturally denote by G_+^*. We observe that $a_{ij}(u) = G^*_{+,ij}(z+u)$. The matrix $||G_+^*||$ of transition probabilities is stochastic due to the positive drift, and it has one irreducibility class by the same argument as that following Theorem 2.2.1. We let $\nu^* = (\nu_i^*)_{i\in E}$ denote the (unique) left invariant probability vector of $||G_+^*||$ and we let

$$\mu^* = \sum_{i,j} \nu_i^* \sum_{u\geq 1} u G^*_{+,ij}(u). \tag{2.15}$$

The Markov key renewal theorem, cf. (2.1) in Example 2.1.1, then gives that

$$A_{ij}(u) \to \frac{1}{\mu^*} \sum_k \nu_k^* \sum_{u\geq 0} a_{kj}(u) = \frac{1}{\mu^*} \sum_k \nu_k^* \sum_{u\geq 0} G^*_{+,kj}(z+u) \tag{2.16}$$

for $u \to \infty$. Hence we conclude that $(J_{\tau(u)}, S_{\tau(u)-u}) \xrightarrow{\mathcal{D}} (J_\infty, B_\infty)$, where

$$\mathbb{P}(B_\infty = z, J_\infty = j) = \frac{1}{\mu^*} \sum_k \nu_k^* \sum_{u\geq z} G^*_{+,kj}(u) = \frac{1}{\mu^*} \sum_k \nu_k^* \overline{G^*_{+,kj}}(z-1),$$

using the notation $\overline{G^*_{+,kj}}(z) = \sum_{u>z} G^*_{+,kj}(u)$ to denote the tail of the distribution function. Especially

$$\mathbb{E}^*_i\left(\frac{1}{r^*_{J_{\tau(u)}}}\exp(-\theta^*(S_{\tau(u)}-u))\right) \to c = \frac{1}{\mu^*}\sum_{\substack{z\geq 1\\ j,k\in E}}\frac{\nu^*_k}{r^*_j}\exp(-\theta^* z)\overline{G^*_{+,kj}}(z-1) \quad (2.17)$$

for $u \to \infty$.

For the remaining part of the result, let $f : E \times \mathbb{N} \to \mathbb{R}$ be any bounded function and put $h_i(u) = \mathbb{E}^*_i(f(J_{\tau(u)}, S_{\tau(u)-u}))$ so that $h_i(u) \to h(\infty)$ for $u \to \infty$ independent of $i \in E$. Let $u' = \lfloor u/2 \rfloor$ in which case we get

$$\begin{aligned}\mathbb{E}^*_i(f(J_{\tau(u)}, S_{\tau(u)-u}) \mid \mathcal{F}_{\tau(u')}) &= h_{J_{\tau(u')}}(u - S_{\tau(u')})1(S_{\tau(u')} \leq u)\\ &\quad + f(J_{\tau(u')}, S_{\tau(u')-u})1(S_{\tau(u')} > u).\end{aligned}$$

Clearly for $u \to \infty$ we have that $1(S_{\tau(u')} > u) \leq 1(S_{\tau(u')} - u' > u') \xrightarrow{\mathbb{P}^*_i} 0$ as $S_{\tau(u')} - u' \xrightarrow{\mathcal{D}} B_\infty$, and likewise $u - S_{\tau(u')} = u - u' + (u' - S_{\tau(u')}) \xrightarrow{\mathbb{P}^*_i} \infty$. Hence

$$\mathbb{E}^*_i(f(J_{\tau(u)}, S_{\tau(u)-u}) \mid \mathcal{F}_{\tau(u')}) \xrightarrow{\mathbb{P}^*_i} h(\infty) = \mathbb{E}(f(J_\infty, B_\infty)).$$

Observing that $\{\tau_- > \tau(u)\} \searrow \{\tau_- = \infty\}$ and $\{\tau_- > \tau(u')\} \searrow \{\tau_- = \infty\}$ a.s. for $u \to \infty$ yields, using the boundedness of f, that

$$\begin{aligned}\lim_{u\to\infty}\mathbb{E}^*_i(f(J_{\tau(u)}, S_{\tau(u)-u}); \tau_- > \tau(u)) &= \lim_{u\to\infty}\mathbb{E}^*_i(f(J_{\tau(u)}, S_{\tau(u)-u}); \tau_- > \tau(u'))\\ &= \lim_{u\to\infty}\mathbb{E}^*_i(\mathbb{E}^*_i(f(J_{\tau(u)}, S_{\tau(u)-u})\ |\mathcal{F}_{\tau(u')}); \tau_- > \tau(u'))\\ &= \mathbb{E}(f(J_\infty, B_\infty))\mathbb{P}^*_i(\tau_- = \infty).\end{aligned}$$

By the exponential change of measure (2.13), this completes the proof. □

Remark 2.4.2 *The global maximum,*

$$\mathcal{M} = \max_n S_n,$$

is finite due to the negative drift, and it is of some interest in itself. Exactly as in the proof above, substituting $(\tau_- > \tau(u))$ *with* $(\tau(u) < \infty)$*, we find that*

$$\mathbb{P}_i(\mathcal{M} > u) \sim c\, r^*_i \exp(-\theta^* u) \quad (2.18)$$

for $u \to \infty$*. The interest in* $\mathcal{M}$ *stems for instance from the fact that the invariant distribution of the reflected MAP* $(J_n, T_n)_{n\geq 0}$ *coincides with the distribution of* $(\overleftarrow{J_0}, \overleftarrow{\mathcal{M}})$ *where* $\overleftarrow{\mathcal{M}} = \max_n \overleftarrow{S}_n$ *and* $\mathcal{D}(\overleftarrow{J_0}) = \pi$*, cf. Proposition XI.2.11 in Asmussen (2003).*

As a consequence of Theorem 2.4.1 we obtain:

Theorem 2.4.3 *With ν the left invariant probability vector for $||G_-||$ and*

$$K^* = \frac{c}{\mathbb{E}_\nu(\tau_-)}\ \nu C = \frac{c}{\mathbb{E}_\nu(\tau_-)} \sum_i \nu_i c_i \tag{2.19}$$

where c and $C = (c_i)_{i\in E}$ are the constants from Theorem 2.4.1 we have that

$$\mathbb{P}_{i_0}(\mathcal{M}_\sigma > u) \sim \mathbb{E}_{i_0}(\sigma) K^* \exp(-\theta^* u) \tag{2.20}$$

for $u \to \infty$.

For the proof we need the following lemma.

Lemma 2.4.4 *Let Γ be the matrix given by*

$$\Gamma_{ij} = \begin{cases} ||G_-||_{ij} & \text{if } j \neq i_0 \\ 0 & \text{if } j = i_0 \end{cases},$$

then $I - \Gamma$ is invertible and

$$(I-\Gamma)^{-1}_{i_0 j} = \nu_j \frac{\mathbb{E}_{i_0}(\sigma)}{\mathbb{E}_\nu(\tau_-)},$$

where $\nu = (\nu_j)_{j\in E}$ is the left invariant probability vector for $||G_-||$.

Proof: Irreducibility of $||G_-||$ implies that $I - \Gamma$ is invertible with

$$(I-\Gamma)^{-1} = \sum_{n\geq 0} \Gamma^n.$$

The value of Γ^n_{ij} is the probability that the Markov chain $(J_{\tau_-(n)})_{n\geq 0}$ jumps from state i to state $j \neq i_0$ in n-steps avoiding state i_0, and is known as a taboo probability. It can be verified that $((I-\Gamma)^{-1}_{i_0 j})_{j\in E}$ is in fact an (unnormalised) left invariant vector for $||G_-||$, which is hence proportional to ν. Observing that

$$\sum_{j\in E} \Gamma^n_{i_0 j} = \mathbb{P}_{i_0}(\widetilde{\sigma} > n)$$

yields that

$$\sum_{j\in E} (I-\Gamma)^{-1}_{i_0 j} = \mathbb{E}_{i_0}(\widetilde{\sigma}) = \frac{\mathbb{E}_{i_0}(\sigma)}{\mathbb{E}_\nu(\tau_-)},$$

where the last equality follows from Wald's identity as stated in Remark 2.3.3 for the MAP $(\sum_{k=1}^{n} \tau_-(k), J_{\tau_-(n)})_{n\geq 0}$. □

Proof of Theorem 2.4.3 Using the strong Markov property for $(J_n, T_n)_{n\geq 0}$ at time τ_- and that the processes $(T_n)_{n\geq 0}$ and $(S_n)_{n\geq 0}$ coincide up to time $\tau_- - 1$ we have that

$$\begin{aligned}\mathbb{P}_i(\mathcal{M}_\sigma > u) &= \mathbb{P}_i(\mathcal{M}_\sigma > u, \mathcal{M}_{\tau_-} > u) + \sum_{j\neq i_0} \mathbb{P}_i(\mathcal{M}_\sigma > u, \mathcal{M}_{\tau_-} \leq u, J_{\tau_-} = j) \\ &= \mathbb{P}_i(\mathcal{M}_{\tau_-} > u) + \sum_{j\neq i_0} \mathbb{P}_i(\mathcal{M}_{\tau_-} \leq u, J_{\tau_-} = j)\mathbb{P}_j(\mathcal{M}_\sigma > u).\end{aligned}$$

Put $A_i(u) = \mathbb{P}_i(\mathcal{M}_\sigma > u)$, $B_i(u) = \mathbb{P}_i(\mathcal{M}_{\tau_-} > u)$ and $\Gamma_{ij}(u) = \mathbb{P}_i(\mathcal{M}_{\tau_-} \leq u, J_{\tau_-} = j)$ for $j \neq i_0$ together with $\Gamma_{ii_0}(u) = 0$. Then in vector notation the equations above can be rewritten as

$$(I - \Gamma(u))A(u) = B(u).$$

Observe that $\Gamma(u) \to \Gamma$ defined in Lemma 2.4.4. Since the set of invertible matrices is open, $I - \Gamma(u)$ is invertible for u large enough and we obtain that

$$A(u) \sim (I - \Gamma)^{-1} B(u)$$

for $u \to \infty$. Using Lemma 2.4.4 and Theorem 2.4.1 this implies that

$$\mathbb{P}_{i_0}(\mathcal{M}_\sigma > u) \sim \sum_j (I - \Gamma)^{-1}_{i_0 j} B(u)_j \sim \mathbb{E}_{i_0}(\sigma) K^* \exp(-\theta^* u).$$

□

Remark 2.4.5 *By Remark 2.4.2 we have the following explicit expression*

$$\mathbb{E}_{i_0}(\sigma) = \frac{1}{\pi_{i_0}\mathbb{P}_{i_0}(\overleftarrow{\mathcal{M}} = 0)} = \frac{1}{\pi_{i_0}\left(1 - \sum_{j\in E} ||\overleftarrow{G}_{+,i_0 j}||\right)}$$

for the expected regeneration time appearing above.

Computing K^* boils down to computing c and the vector C from Theorem 2.4.1 together with $\mathbb{E}_\nu(\tau_-)$. For this last mean value, Wald's identity for MAPs, cf. Remark 2.3.3, may again be useful, giving that

$$\mathbb{E}_\nu(\tau_-) = \frac{\mathbb{E}_\nu(S_{\tau_-})}{\mu} = \frac{1}{\mu} \sum_{\substack{u\leq 0 \\ ij}} \nu_i u G_{-,ij}(u) = \frac{1}{\mu} \nu \left(\sum_{u\leq 0} u G_-(u) \right) \mathbb{1}$$

It turns out that C and c, and hence K^*, are computable from θ^*, G_+ and G_-. This is discussed further in Section 2.5 together with some appropriate algorithms for computing G_+ and G_-.

As argued, once the tail behaviour for the distribution of $\mathcal{M}_\sigma$ has been established, the asymptotic behaviour of $\mathcal{M}_n$ for $n \to \infty$ can also be dealt with. The following lemma will serve as a *key lemma* for the results derived in Chapter 5 and 6.

Lemma 2.4.6 *Suppose that $n(u)$ is a sequence of integers satisfying*

$$\frac{u}{n(u)} \to 0 \quad \text{and} \quad n(u)\exp(-\theta^* u) \to 0 \tag{2.21}$$

for $u \to \infty$. Then

$$\mathbb{P}_i(\mathcal{M}_{n(u)} > u) \sim n(u)K^* \exp(-\theta^* u) \tag{2.22}$$

for $u \to \infty$.

Proof: Let $G(u) = \mathbb{P}_{i_0}(\mathcal{M}_\sigma \leq u)$ be the distribution function for $\mathcal{M}_\sigma$. Introduce the sequence $(\sigma(m))_{m\geq 0}$ of stopping times by $\sigma(0) = 0$ and

$$\sigma(m) = \inf\{k > \sigma(m-1) \mid T_k = 0,\ J_k = i_0\}.$$

Then

$$\mathcal{M}_{\sigma(m)} = \max_{0\leq k\leq \sigma(m)} T_k = \max_{1\leq k\leq m} \max_{\sigma(k-1)\leq r\leq \sigma(k)} T_r,$$

which by regeneration is a max of m *iid* variables. Thus $\mathbb{P}_{i_0}(\mathcal{M}_{\sigma(m)} > u) = 1-G(u)^m$.

Define, for $\delta > 0$ fixed, the sequences $m_-(u) = \gamma_-(u)n(u)/\mathbb{E}_{i_0}(\sigma)$ and $m_+(u) = \gamma_+(u)n(u)/\mathbb{E}_{i_0}(\sigma)$ with $\gamma_-(u) \leq 1-\delta$ and $\gamma_+(u) \geq 1+\delta$ chosen maximally and minimally such that $m_-(u), m_+(u) \in \mathbb{N}$. In particular this implies that $\gamma_-(u) \to 1-\delta$ and $\gamma_+(u) \to 1+\delta$ for $u \to \infty$. Then since $(1-G(u))\exp(\theta^* u) \to \mathbb{E}_{i_0}(\sigma)K^*$ we have, using that $n(u)\exp(\theta^* u) \to 0$, that

$$\begin{aligned}(1-G(u)^{m_-(u)})\frac{\exp(\theta^* u)}{n(u)} &\to (1-\delta)K^* \quad \text{and}\\ (1-G(u)^{m_+(u)})\frac{\exp(\theta^* u)}{n(u)} &\to (1+\delta)K^*.\end{aligned}$$

Since $\mathcal{M}_n$ is clearly increasing, we have the following inequalities

$$\begin{aligned}\mathbb{P}_{i_0}(\mathcal{M}_{\sigma(m_-(u))} > u) - \mathbb{P}_{i_0}(\sigma(m_-(u)) > n(u)) &\leq \mathbb{P}_{i_0}(\mathcal{M}_{n(u)} > u)\\ &\leq \mathbb{P}_{i_0}(\mathcal{M}_{\sigma(m_+(u))} > u) + \mathbb{P}_{i_0}(\sigma(m_+(u)) < n(u)).\end{aligned} \tag{2.23}$$

The stopping time $\sigma(m) = \sum_{k=1}^{m} \sigma(k) - \sigma(k-1)$ is a sum of *iid* variables (under $\mathbb{P}_{i_0}$) with $\mathbb{E}_{i_0}(\exp(\lambda\sigma)) < \infty$ for some $\lambda > 0$, and it follows by large deviation theory that there exists a $\lambda^* > 0$ such that

$$\mathbb{P}_{i_0}(\sigma(m_-(u)) > n(u)) \;\leq\; \exp(-\lambda^* m_-(u)).$$

This implies that

$$\exp(\theta^* u)\mathbb{P}_{i_0}(\sigma(m_-(u)) > n(u)) \to 0$$

by $u/n(u) \to 0$. And similarly $\exp(\theta^* u)\mathbb{P}_{i_0}(\sigma(m_+(u)) > n(u)) \to 0$, which together with (2.23) imply that

$$\begin{aligned}(1-\delta)K^* &\leq \liminf_{u\to\infty} \mathbb{P}_{i_0}(\mathcal{M}_{n(u)} > u)\frac{\exp(\theta^* u)}{n(u)} \\ &\leq \limsup_{u\to\infty} \mathbb{P}_{i_0}(\mathcal{M}_{n(u)} > u)\frac{\exp(\theta^* u)}{n(u)} \leq (1+\delta)K^*.\end{aligned}$$

Letting $\delta \to 0$ gives (2.22) for $i = i_0$ and otherwise the estimate

$$\mathbb{P}_i(\mathcal{M}_\sigma > u) \leq K\exp(-\theta u)$$

ensures that (2.22) also holds for $i \neq i_0$. □

By quite similar methods, assuming instead that $n(u) \sim \exp(\theta^* u)$, one can derive the asymptotic distribution of $\mathcal{M}_n$.

Theorem 2.4.7 *For $z \in \mathbb{R}$ it holds that*

$$\mathbb{P}_i(\theta^*\mathcal{M}_n - \log(K^* n) \leq z) - \exp(-\exp(-z + z_n)) \to 0 \tag{2.24}$$

for $n \to \infty$ where $z_n \in [0, \theta^*)$ is defined by $(\log(K^* n) + z - z_n)/\theta^* \in \mathbb{Z}$.

Basically (2.24) tells us that $\theta^*\mathcal{M}_n - \log(K^* n)$ asymptotically follows an extreme value distribution of type I – the Gumbel distribution – up to the effect of z_n. This z_n-effect comes from the lattice assumption and is a notorious nuisance in formulating theorems but can often be ignored in practice. From the fact that $0 \leq z_n < \theta^*$ we can easily derive upper and lower bounds for $\mathbb{P}_i(\theta^*\mathcal{M}_n - \log(K^* n) \leq z)$ but for many purposes one can simply ignore the z_n-correction.

We skip a direct proof and refer to Section VI.4 in Asmussen (2003), which contains an argument proving this theorem together with several related results. We will in Section 2.6 derive a Poisson approximation of the number of excess events, from which the theorem is an easy consequence anyway.

2.5 Computation of constants

To use the asymptotic theory we will need to be able to compute the constants θ^* and K^*. Computing θ^* numerically is quite easy if we can compute $\Phi(\theta)_{ij} = \mathbb{E}_i(\exp(\theta X_1); J_1 = j)$, because then we can compute numerically the spectral radius $\varphi(\theta)$, and since φ is a convex function the numerical solution of $\varphi(\theta) = 1$ ($\psi(\theta) = 0$) is easy. The computation of K^* is more involved, and we need to compute the vector C and the constant c from Theorem 2.4.1. To this end we show below that, besides θ^*, we 'just' need to compute the matrices G_+ and G_- . We discuss two approaches in this section for computing G_+ and G_-, an iterative approach and a spectral approach.

Lemma 2.5.1 *With ν^* the left invariant probability vector of $||G^*_+||$ and μ^* defined by (2.15) we have that*

$$c = \frac{1}{\mu^*(\exp(\theta^*) - 1)} \frac{\nu^*}{r^*}(I - ||G_+||)\mathbb{1} \tag{2.25}$$

with $\nu^*/r^* = (\nu^*_i/r^*_i)_{i\in E}$. Furthermore,

$$C = (c_i)_{i\in E} = (I - \sum_u \exp(\theta^* u)G_-(u))r^*. \tag{2.26}$$

Proof: Using the exponential change of measure

$$\begin{aligned} G^*_{+,kj}(u) &= \mathbb{P}^*_k(S_{\tau_+} = u, J_{\tau_+} = j) \\ &= \frac{r^*_j}{r^*_k} \exp(\theta^* u)\mathbb{P}_k(S_{\tau_+} = u, J_{\tau_+} = j, \tau_+ < \infty) \\ &= \frac{r^*_j}{r^*_k} \exp(\theta^* u)G_{+,kj}(u). \end{aligned} \tag{2.27}$$

Interchanging the summation order in (2.17), this equality yields

$$\begin{aligned} \sum_{z\geq 1} \exp(-\theta^* z)\overline{G^*_{+,kj}}(z-1) &= \sum_{u\geq 1} G^*_{+,kj}(u) \sum_{z=1}^{u} \exp(-\theta^* z) \\ &= \frac{r^*_j}{r^*_k} \frac{1}{\exp(\theta) - 1} \sum_{u\geq 1} (1 - \exp(-\theta^* u)) \exp(\theta^* u)G_{+,kj}(u) \\ &= \frac{1}{\exp(\theta) - 1} \left(||G^*_{+,kj}|| - \frac{r^*_j}{r^*_k} ||G_{+,kj}|| \right). \end{aligned}$$

Using this, equation (2.17) and the fact that ν^* is the left invariant probability vector for $||G^*_+||$ give in matrix notation the formula (2.25).

Regarding the computation of C we have directly from the definition of G^*_- that

$$\mathbb{P}^*_i(\tau_- = \infty) = 1 - \sum_j ||G^*_{-,ij}||.$$

The analogues change of measure argument as for G_+ gives that

$$||G^*_{-,ij}|| = \frac{r^*_j}{r^*_i} \sum_u \exp(\theta^* u) G_{-,ij}(u).$$

In matrix notation we obtain (2.26). □

Equations (2.25) and (2.26) are convenient and compact ways to represent c and C in terms of G_+ and G_-. Note that ν^* and μ^* are easily computable once G_+ is known using (2.27).

We will assume throughout the rest of this section that the distributions H_{ij} besides being concentrated on the integer lattice also have finite support. Hence we will assume that they are concentrated on $\{-m, \ldots, n\}$.

Recall that F is the matrix with entries $F_{ij} = H_{ij}P_{ij}$ and that $F_{ij}(u) = H_{ij}(u)P_{ij}$. If we want to compute $G_{+,ij}(u)$ – the probability that the MAP jumps to (j, u) at the first ascending ladder time given $J_0 = i$ – we can decompose the sample path of the MAP according to the first jump. It can either jump directly to $u > 0$, which is thus the ascending ladder height, or it can jump to $z \leq 0$ from where it successively has to make ascending ladder jumps $u_1, \ldots, u_r$ fulfilling that $u_1 + \cdots + u_r = u - z$ and $u_1 + \cdots + u_{r-1} \leq -z$. In matrix notation this amounts to the following equation

$$G_+(u) = F(u) + F(0)G_+(u) + \sum_{z=-1}^{-m} F(z) \sum_{\substack{u_1+\ldots+u_{r-1}\leq -z \\ u_1+\ldots+u_r = u-z \\ 1\leq u_i \leq n}} G_+(u_1)G_+(u_2)\cdots G_+(u_r)$$

for $1 \leq u \leq n$.

Introduce the map ρ_+ defined by

$$\rho_+(G)(u) = F(u) + F(0)G(u) + \sum_{z=-1}^{-m} F(z) \sum_{\substack{u_1+\ldots+u_{r-1}\leq -z \\ u_1+\ldots+u_r = u-z \\ 1\leq u_i \leq n}} G(u_1)G(u_2)\cdots G(u_r).$$

Then G_+ is a fixed point for the map ρ_+. This can be employed in an iterative algorithm for computing G_+.

Theorem 2.5.2 *The matrix G_+ is the coordinatewise minimal fixed point for ρ_+, and taking $G^{(0)} = 0$ the sequence $(G^{(k)})_{k\geq 0}$ defined recursively for $k \geq 1$ by*

$$G^{(k)} = \rho_+(G^{(k-1)}) \tag{2.28}$$

converges coordinatewise monotonely to G_+.

Proof: Assume that G is a matrix of positive measures satisfying $\rho_+(G) = G$. Then $G \geq 0 = G^{(0)}$ (coordinatewise) and also $G^{(1)}(u) = F(u) \geq 0 = G^{(0)}(u)$ for $1 \leq u \leq n$. Hence for $k \geq 1$ we get by induction that

$$\begin{aligned} G^{(k+1)}(u) - G^{(k)}(u) &= F(0)(G^{(k)}(u) - G^{(k-1)}(u)) \\ &+ \sum_{z=-1}^{-m} F(z) \sum_{\substack{u_1+\ldots+u_{r-1}\leq -z \\ u_1+\ldots+u_r = u-z \\ 1\leq u_i \leq n}} G^{(k)}(u_1)\cdots G^{(k)}(u_r) - G^{(k-1)}(u_1)\cdots G^{(k-1)}(u_r) \;\geq\; 0, \end{aligned}$$

since matrix multiplication preserves coordinatewise ordering. Likewise by induction

$$\begin{aligned} G(u) - G^{(k)}(u) &= F(0)(G(u) - G^{(k-1)}(u)) \\ &+ \sum_{z=-1}^{-m} F(z) \sum_{\substack{u_1+\ldots+u_{r-1}\leq -z \\ u_1+\ldots+u_r = u-z \\ 1\leq u_i \leq n}} G(u_1)\cdots G(u_r) - G^{(k-1)}(u_1)\cdots G^{(k-1)}(u_r) \;\geq\; 0. \end{aligned}$$

This implies that $G^{(k)} \nearrow L$ for some matrix L with $L \leq G$ and by continuity this limit must also satisfy $\rho_+(L) = L$. Hence L is the minimal fixed point for ρ_+, and in particular we have that $G_+ \geq L$.

To show that G_+ is indeed equal to the minimal fixed point, introduce the variable

$$\gamma = \sum_{l=1}^{\tau_+} 1(Z_l \leq 0),$$

which count the number of descends before the first ascending ladder time τ_+. Introduce also the matrix

$$\widetilde{G}_{ij}^{(n)}(u) = \mathbb{P}_i(J_{\tau_+} = j, S_{\tau_+} = u, \gamma \leq n).$$

Thus $\widetilde{G}^{(n)}$ is the matrix of ascending ladder height distributions under the restriction that up to time τ_+ only n descends may occur in total. Dividing according to the value of Z_1 gives

$$\widetilde{G}_{ij}^{(n+1)}(u) = F_{ij}(u) + \sum_k F_{ik}(0) G_{kj}^{(n)}(u) + \sum_{z=-1}^{-m} \mathbb{P}_i(J_{\tau_+} = j, S_{\tau_+} = u, \gamma \leq n+1, Z_1 = z)$$

Since $Z_1 = z < 0$ in the last term above, this is one descend and the MAP has to make ascends from z of total size $u - z$ (with the last ascending ladder height being at least u) under the restriction that at most n descends happen in-between.

Clearly this probability is smaller than if we allow n descends to happen between each successive ascending ladder epoch. Hence

$$\widetilde{G}^{(n+1)}(u) \leq \rho_+(\widetilde{G}^{(n)})(u).$$

But we can also observe that $\widetilde{G}^{(0)}(u) = F(u) = G^{(1)}(u)$, hence by induction

$$\widetilde{G}^{(n+1)}(u) \leq \rho_+(\widetilde{G}^{(n)})(u) \leq \rho_+(G^{(n+1)})(u) = G^{(n+2)}(u).$$

For $n \to \infty$ we clearly have that $\widetilde{G}^{(n+1)}(u) \nearrow G_+(u)$ and we obtain that $G_+ \leq L$. This proves that G_+ is the minimal fixed point. □

Defining

$$\rho_-(G)(u) = F(u) + \sum_{z=1}^{n} F(z) \sum_{\substack{u_1+\ldots+u_{r-1}\geq -z \\ u_1+\ldots+u_r = u-z \\ -m\leq u_i\leq 0}} G(u_1)G(u_2)\cdots G(u_r). \tag{2.29}$$

for $-m \leq u \leq 0$ we obtain the completely analogous result; G_- is the minimal fixed point for ρ_- and applying ρ_- iteratively starting with $G^{(0)} = 0$ yields coordinatewise and monotone convergence towards G_-.

The other approach to the computation of G_+ and G_- relies on the Wiener-Hopf factorisation identity. Since the matrix measures we will consider are concentrated on $\{-m, \ldots, n\}$, the coordinatewise generating functions are well defined on $\mathbb{C}\backslash\{0\}$. We let $\widehat{F}[s]$ denote the matrix of generating functions for F evaluated at s.

Theorem 2.5.3 *If $s_1, \ldots, s_q$ are the roots with $|s_k| > 1$ of*

$$\det(I - \widehat{F}[s]) = 0 \tag{2.30}$$

and $a_1, \ldots, a_q$ the corresponding right eigenvectors, i.e. $a_k = \widehat{F}[s_k]a_k$, the matrix G_+ fulfills the equations

$$a_k = s_k G_+(1)a_k + s_k^2 G_+(2)a_k + \cdots + s_k^n G_+(n)a_k \tag{2.31}$$

for $k = 1, \ldots, q$. Similarly, if $t_1, \ldots, t_r$ are the roots of (2.30) with $|t_k| \leq 1$ and $b_1, \ldots, b_r$ are the corresponding right eigenvectors, then G_- fulfills

$$b_k = G_-(0)b_k + t_k^{-1}G_-(-1)b_k + \cdots + t_k^{-m}G_-(-m)b_k \tag{2.32}$$

for $k = 1, \ldots, r$.

Proof: Let $\widehat{G}_-[s]$ and $\widehat{G}_+[s]$ denote the matrices of generating functions for the descending and ascending ladder height distributions. Notice that $\widehat{G}_-[s]$ is a matrix of polynomials in $1/s$ and $\widehat{G}_+[s]$ a matrix of polynomials in s whereas $\widehat{F}[s]$ contains mixed terms. The Wiener-Hopf identity (2.2) implies that

$$(I - \widehat{F}[s]) = (I - \widehat{{}^{\#}G_+}[s])(I - \widehat{G}_-[s]) = (I - \widehat{{}^{\#}G_-}[s])(I - \widehat{G}_+[s]).$$

If $|s| > 1$ then $\mathrm{spr}(\widehat{{}^{\#}G_-}[s]) < 1$, hence a right eigenvector for $\widehat{F}[s]$ with eigenvalue 1 must also be a right eigenvector with eigenvalue 1 for $\widehat{G}_+[s]$. Likewise, for $|s| \leq 1$ we have that $\mathrm{spr}(\widehat{{}^{\#}G_+}[s]) < 1$ (when $|s| = 1$ this is a result of the negative drift assumption), in which case a right eigenvector for $\widehat{F}[s]$ with eigenvalue 1 must be a right eigenvector with eigenvalue 1 for $\widehat{G}_-[s]$. Taking determinants, this observation implies that the roots of

$$\det(I - \widehat{F}[s]) = 0$$

coincides for $|s| > 0$ with the roots of $\det(I - \widehat{G}_+[s]) = 0$ and for $|s| \leq 1$ with the roots of $\det(I - \widehat{G}_-[s]) = 0$. And the right eigenvectors corresponding to the roots must also coincide. □

Since we can compute the roots and eigenvectors from F, we can hope that there are sufficiently many roots so that G_+ and G_- can be identified uniquely. The matrix G_+ contains $n|E|^2$ unknowns, so without other knowledge we will need at least $n|E|$ roots to have sufficiently many equations given by (2.31) to solve for G_+. Clearly $\det(I - \widehat{G}_+[s])$ can at most have degree $n|E|$, hence there can at most be $n|E|$ roots. But what is the degree of polynomials like $\det(I - \widehat{G}_+[s])$, and hence the number of roots obtainable? Consider a general matrix polynomial $A[s]$. Write the polynomial as

$$A[s] = A(0) + A(1)s + \cdots + A(n)s^n$$

with $A(0), \ldots, A(n)$ the coefficient matrices. Letting $\mathcal{S}$ denote the set of permutations on E, the formula

$$\det(A[s]) = \sum_{\sigma \in \mathcal{S}} (-1)^{|\sigma|} \prod_{i \in E} A[s]_{i\sigma(i)}$$

for computing the determinant implies that the term of degree $n|E|$ have coefficient $\det(A(n))$, which is thus $\neq 0$ if and only if $A(n)$ is invertible.

An insufficient number of equations may occur for other reasons, also in the random walk case, namely if some roots have multiplicity > 1. This is most likely only a

theoretical problem that doesn't occur in practical applications, but it is nevertheless noteworthy.

A practical solution to the problem of an insufficient number of equations is to simply find the roots of $\det(I - \hat{F}[s]) = 0$, then by other means figuring out which entries in G_+ and G_- are equal to 0, and then finally hope that we have a sufficient number of equations to solve for the remaining entries $\neq 0$. For the practical applications discussed in this thesis in Chapter 7, the recursive algorithm given above was implemented for the computation of G_+ and G_- and hence K^*.

2.6 A Poisson approximation

The last topic of this chapter is a Poisson approximation result for the sum of possibly dependent Bernoulli variables. We will illustrate the use of this result by studying the number of exceedances over a threshold for a reflected MAP, but before doing so, we make a brief digression to certain relations between mixing coefficient. The application is supposed to be an appetiser of the main results and techniques of proof in Chapter 5 and Chapter 6.

Assume that I is some (finite) index set and let $(V_a)_{a\in I}$ be a collection of Bernoulli random variables. For each $a \in I$ we assume that a subset $B_a \subseteq I$ is given. Furthermore, for $a \in I$ let

$$\mathcal{F}_a = \sigma(V_b \mid b \notin B_a)$$

be the σ-algebra generated by the variables *not* in B_a.

Rephrasing Theorem 1 in Arratia et al. (1989) gives:

Theorem 2.6.1 *With* $\mathrm{Poi}(\lambda)$ *the Poisson distribution with mean* λ *for* $\lambda > 0$*, the inequality*

$$\left|\left|\mathcal{D}\left(\sum_{a\in I} V_a\right) - \mathrm{Poi}\left(\sum_{a\in I}\mathbb{E}(V_a)\right)\right|\right| \leq 2\beta \tag{2.33}$$

holds with

$$\beta = \sum_{a\in I, b\in B_a}\mathbb{E}(V_a)\mathbb{E}(V_b) + \sum_{a\in I, b\in B_a, b\neq a}\mathbb{E}(V_aV_b) + \sum_{a\in I}\mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)|. \tag{2.34}$$

Note that since $||\mathrm{Poi}(\lambda_1) - \mathrm{Poi}(\lambda_2)|| \leq |\lambda_1 - \lambda_2|$, we can also obtain a Poisson approximation using an approximation of the mean value of $\sum_a \mathbb{E}(V_a)$. This is useful if we can not compute $\sum_a \mathbb{E}(V_a)$ but have a sufficiently good approximation.

The theorem is credited to Arratia et al. (1989), who use the Chen-Stein-method. In their own words, the essential ingredients for proving Theorem 1 in Arratia et al. (1989), and hence Theorem 2.6.1 above, are contained in the original paper by Chen (1975). The theorem holds as stated for all choices of sets B_a, $a \in I$, but it works best (gives best bounds), if the sets are chosen with some care. In several cases, the sets can be chosen such that the variables V_b for $b \notin B_a$ are independent of V_a, in which case the last term in β disappears. In this case it is common to call the set B_a the *neighbourhood of dependence* of the variable V_a. More generally, they must be chosen so that there is a suitable tradeoff between the first two terms and the last term in β. Since $\mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)|$ measures dependence between the variable V_a and the σ-algebra $\mathcal{F}_a$, and since we want it to be sufficiently small, it is reasonable in general to call B_a the *neighbourhood of strong dependence* of V_a.

2.6.1 A note on mixing

To apply the theorem above – when the last term in β does not vanish – we need to somehow bound variables of the form $\mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)|$. We discuss in this section a few useful results for giving bounds on such variables.

For two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ (with $(\Omega, \mathcal{H}, \mathbb{P})$ a probability field and $\mathcal{F}, \mathcal{G} \subseteq \mathcal{H}$) we define the α-mixing measure of dependence as

$$\alpha(\mathcal{F}, \mathcal{G}) = \sup_{A \in \mathcal{F}, B \in \mathcal{G}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|.$$

The first lemma relates α-mixing measures to mean values of the desired form $\mathbb{E}|\mathbb{E}(\eta|\mathcal{F}) - \mathbb{E}(\eta)|$.

Lemma 2.6.2 *Let $\mathcal{F}$ and $\mathcal{G}$ be σ-algebras and let $A \in \mathcal{G}$. With $\eta = 1(A)$*

$$\mathbb{E}|\mathbb{E}(\eta|\mathcal{F}) - \mathbb{E}(\eta)| \leq 2\alpha(\mathcal{F}, \mathcal{G}). \tag{2.35}$$

Proof: With $B = (\mathbb{E}(\eta|\mathcal{F}) \geq \mathbb{E}(\eta)) \in \mathcal{F}$ and $\xi = 1(B)$ we see that

$$\begin{aligned} \mathbb{E}|\mathbb{E}(\eta|\mathcal{F}) - \mathbb{E}(\eta)| &= \mathbb{E}(\xi(\mathbb{E}(\eta|\mathcal{F}) - \mathbb{E}(\eta))) - \mathbb{E}((1-\xi)(\mathbb{E}(\eta|\mathcal{F}) - \mathbb{E}(\eta))) \\ &= 2(\mathbb{E}(\xi\eta) - \mathbb{E}(\xi)\mathbb{E}(\eta)) \\ &= 2(\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)) \ \leq \ 2\alpha(\mathcal{F}, \mathcal{G}). \end{aligned}$$

□

The β-mixing measure of dependence between the σ-algebras $\mathcal{F}$ and $\mathcal{G}$ is defined as

$$\beta(\mathcal{F}, \mathcal{G}) = \mathbb{E}(\sup_{A \in \mathcal{F}} |\mathbb{P}(A|\mathcal{G}) - \mathbb{P}(A)|).$$

Lemma 2.6.3 *If $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{G}$ are σ-algebras it holds, with $\mathcal{F}_1 \vee \mathcal{F}_2$ denote the smallest σ-algebra containing $\mathcal{F}_1$ and $\mathcal{F}_2$, that*

$$\alpha(\mathcal{F}_1 \vee \mathcal{F}_2, \mathcal{G}) \leq \alpha(\mathcal{F}_2, \mathcal{G}) + \beta(\mathcal{F}_2 \vee \mathcal{G}, \mathcal{F}_1) + \beta(\mathcal{F}_2, \mathcal{F}_1). \tag{2.36}$$

Proof: Introduce the sets

$$\mathscr{B} = \left\{ \bigcup_{j=1}^{n} A_{1j} \cap A_{2j} \mid A_{1j} \in \mathcal{F}_1,\ A_{2j} \in \mathcal{F}_2,\ n \geq 1 \right\}$$

and, with $c = \alpha(\mathcal{F}_2, \mathcal{G}) + \beta(\mathcal{F}_2 \vee \mathcal{G}, \mathcal{F}_1) + \beta(\mathcal{F}_2, \mathcal{F}_1)$,

$$\mathscr{A} = \{A \in \mathcal{F}_1 \vee \mathcal{F}_2 \mid \forall B \in \mathcal{G} : |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \leq c\}.$$

We prove that $\mathscr{B}$ is an algebra and that $\mathscr{A}$ is a monotone class containing $\mathscr{B}$.

We see that indeed $\emptyset, \Omega \in \mathscr{B}$ and that $\mathscr{B}$ is closed under finite unions. If $A = \cup_{j=1}^{n} A_{1j} \cap A_{2j}$ we have that

$$A^c = \bigcap_{j=1}^{n} A_{1j}^c \cup A_{2j}^c = \bigcup_{f} \bigcap_{j=1}^{n} A_{f(j)j}^c,$$

where the last union is taken over all functions $f : \{1, \ldots, n\} \to \{1, 2\}$. Observe that

$$\bigcap_{j=1}^{n} A_{f(j)j}^c = \bigcap_{j:f(j)=1} A_{1j}^c \cap \bigcap_{j:f(j)=2} A_{2j}^c,$$

so defining $\tilde{A}_{if} = \cap_{j:f(j)=i} A_{ij}^c \in \mathcal{F}_i$ for $i = 1, 2$ we see that

$$A^c = \bigcup_{f} \tilde{A}_{1f} \cap \tilde{A}_{2f} \in \mathscr{B},$$

and $\mathscr{B}$ is an algebra.

Using continuity of the probability measure $\mathbb{P}$ we see that $\mathscr{A}$ is a monotone class.

Finally, we show that $\mathscr{B} \subseteq \mathscr{A}$. Let $A = \cup_{j=1}^{n} A_{1j} \cap A_{2j} \in \mathscr{B}$ where $A_{1j} \in \mathcal{F}_1$ and $A_{2j} \in \mathcal{F}_2$. We can assume w.l.o.g. that $A_{11}, \ldots, A_{1n}$ are disjoint. We get that for

$B \in \mathcal{G}$

$$\begin{aligned}
|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| &= \left| \sum_{j=1}^{n} \mathbb{P}(A_{1j} \cap A_{2j} \cap B) - \mathbb{P}(A_{1j} \cap A_{2j})\mathbb{P}(B) \right| \\
&\leq \left| \sum_{j=1}^{n} \mathbb{P}(A_{1j} \cap A_{2j} \cap B) - \mathbb{P}(A_{1j})\mathbb{P}(A_{2j} \cap B) \right| \\
&+ \left| \sum_{j=1}^{n} \mathbb{P}(A_{1j})\mathbb{P}(A_{j2} \cap B) - \mathbb{P}(A_{1j})\mathbb{P}(A_{2j})\mathbb{P}(B) \right| \\
&+ \left| \sum_{j=1}^{n} \mathbb{P}(A_{1j} \cap A_{2j})\mathbb{P}(B) - \mathbb{P}(A_{1j})\mathbb{P}(A_{2j})\mathbb{P}(B) \right|.
\end{aligned}$$

For the first term we have that

$$\begin{aligned}
&\left| \sum_{j=1}^{n} \mathbb{P}(A_{1j} \cap A_{2j} \cap B) - \mathbb{P}(A_{1j})\mathbb{P}(A_{2j} \cap B) \right| \\
&\qquad \leq \sum_{j=1}^{n} \mathbb{E}(1_{A_{1j}} |\mathbb{P}(A_{2j} \cap B|\mathcal{F}_{I_1}) - \mathbb{P}(A_{2j} \cap B)|) \leq \beta(\mathcal{F}_2 \vee \mathcal{G}, \mathcal{F}_1).
\end{aligned}$$

For the second term

$$\begin{aligned}
&\left| \sum_{j=1}^{n} \mathbb{P}(A_{1j})\mathbb{P}(A_{j2} \cap B) - \mathbb{P}(A_{1j})\mathbb{P}(A_{2j})\mathbb{P}(B) \right| \\
&\qquad \leq \sum_{j=1}^{n} \mathbb{P}(A_{1j}) |\mathbb{P}(A_{j2} \cap B) - \mathbb{P}(A_{2j})\mathbb{P}(B)| \leq \alpha(\mathcal{F}_2, \mathcal{G}).
\end{aligned}$$

And for the last term we have

$$\begin{aligned}
&\left| \sum_{j=1}^{n} \mathbb{P}(A_{1j} \cap A_{2j})\mathbb{P}(B) - \mathbb{P}(A_{1j})\mathbb{P}(A_{2j})\mathbb{P}(B) \right| \\
&\qquad \leq \mathbb{P}(B) \sum_{j=1}^{n} \mathbb{E}(1_{A_{1j}} |\mathbb{P}(A_{2j}|\mathcal{F}_{I_1}) - \mathbb{P}(A_{2j})|) \leq \beta(\mathcal{F}_2, \mathcal{F}_1).
\end{aligned}$$

This proves that $A \in \mathscr{A}$ and hence $\mathscr{B} \subseteq \mathscr{A}$.

Obviously $\sigma(\mathscr{B}) = \mathcal{F}_1 \vee \mathcal{F}_2$ and we conclude that in fact $\mathscr{A} = \mathcal{F}_1 \vee \mathcal{F}_2$ (Neveu 1965, Proposition I.4.2), and (2.36) holds by the definition of the α-mixing measure of dependence. □

For a stationary stochastic process $(Z_n)_{n\in\mathbb{Z}}$ taking values in some set E, and for a subset $I \subseteq \mathbb{Z}$, we define the corresponding σ-algebra $\mathcal{F}_I = \sigma(Z_n; n \in I)$. The β-mixing coefficient is defined as

$$\beta(n) = \beta(\mathcal{F}_{[n,\infty)}, \mathcal{F}_{(-\infty,0]}) = \mathbb{E} \sup_{A\in\mathcal{F}_{[n,\infty)}} |\mathbb{P}(A|\mathcal{F}_{(-\infty,0]}) - \mathbb{P}(A)|, \tag{2.37}$$

for $n \geq 1$ and the process $(Z_n)_{n\in\mathbb{Z}}$ is called β-mixing, if $\beta(n) \to 0$ for $n \to \infty$. For two subsets $I, J \subseteq \mathbb{Z}$, the distance, $d(I, J)$, between the sets is defined as

$$d(I, J) = \inf_{n\in I, m\in J} |n - m|.$$

We call $I \subseteq \mathbb{Z}$ an *interval* if I is either empty, if $I = \{n, n+1, \ldots, m-1, m\}$ for some $n \leq m \in \mathbb{Z}$, in which case we write $I = [n, m]$, or if $I = \{\ldots, n-1, n\} = (-\infty, n]$ or $I = \{n, n+1, \ldots, \} = [n, \infty)$. If $I, J \subseteq \mathbb{Z}$ are two subsets of integers, we write $I < J$ if $n < m$ for all $n \in I$ and $m \in J$.

Theorem 2.6.4 *Assume that $I_1 < I_2 < \ldots < I_\kappa$ is an increasing sequence of intervals in $\mathbb{Z}$ with $\kappa \geq 2$. With with $I = \cup_{i\ odd} I_i$ and $J = \cup_{i\ even} I_i$ it holds that*

$$\alpha(\mathcal{F}_I, \mathcal{F}_J) \leq (2\kappa - 3)\beta(d(I, J)).$$

Proof: The proof is by induction on κ. For $\kappa = 2$, $I_1 \subseteq (-\infty, m]$ and $I_2 \subseteq [n, \infty)$ with $d(I_1, I_2) = d(I, J) = n - m$. By stationarity we get that

$$\begin{aligned}
\alpha(\mathcal{F}_I, \mathcal{F}_J) &\leq \sup_{A\in\mathcal{F}_{(-\infty,0]}, B\in\mathcal{F}_{[n-m,\infty)}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \\
&\leq \sup_{B\in\mathcal{F}_{[n-m,\infty)}} \mathbb{E}(1_B \sup_{A\in\mathcal{F}_{[n-m,\infty)}} |\mathbb{P}(A|\mathcal{F}_{(-\infty,0]}) - \mathbb{P}(A)|) \\
&\leq \beta(n - m) = \beta(d(I, J)).
\end{aligned}$$

So the result holds for $\kappa = 2$. For the induction step, assume that the result holds for $\kappa - 1 \geq 2$ intervals. With $\mathcal{F}_1 = \mathcal{F}_{I_1}$, $\mathcal{F}_2 = \mathcal{F}_{I\setminus I_1}$ and $\mathcal{G} = \mathcal{F}_J$ we get from Lemma 2.6.3 that

$$\alpha(\mathcal{F}_I, \mathcal{F}_J) \;=\; \alpha(\mathcal{F}_1 \vee \mathcal{F}_2, \mathcal{G}) \;\leq\; \alpha(\mathcal{F}_2, \mathcal{G}) + \beta(\mathcal{F}_2 \vee \mathcal{G}, \mathcal{F}_1) + \beta(\mathcal{F}_1, \mathcal{F}_2)$$

The first term is bounded by $(2(\kappa - 1) - 3)\beta(d(I, J))$ by the induction hypothesis and both of the two other terms are bounded by $\beta(d(I, J))$, and the result follows. □

2.6.2 The counting of exceedances for a MAP

We want to use the Poisson approximation in Theorem 2.6.1 to approximate the number of times the reflected MAP exceeds some level t, but we can not do so directly, since such excesses occur in clumps. Thus we need a way to declump the excesses. This is done by decomposing the reflected MAP into clumps and then count whether or not a clump contains an excursion exceeding t.

The path of the reflected MAP $(T_n)_{n\geq 0}$ may, as described in Section 2.4, be decomposed into regenerative cycles given by the descending ladder epochs. In between, the process makes excursions into the positive halfline $(0,\infty)$. It seems natural to divide the process into clumps with each clump corresponding to such an excursion. We define for $n \geq 1$ and $k \geq 0$ the stopping time

$$e_n(k) = \min(\tau_-(k), n).$$

For $k \geq 1$ and $e_n(k-1) < n$, the k'th excursion *before time* n occur from $e_n(k-1) = \tau_-(k-1)$ to $e_n(k)$. Note that the last excursion before time n may not be complete. The maximum of the k'th excursion when $e_n(k-1) < n$ is defined as

$$\mathcal{M}^k = \max_{e_n(k-1)\leq m<e_n(k)} T_m.$$

For $t > 0$ a given threshold we define for $e_n(k-1) < n$

$$U_k(t) = 1\left(\mathcal{M}^k > t\right),$$

which count whether the excursion occurring between time $e_n(k-1)$ and $e_n(k)$ exceeds level t. We note that with these definitions

$$(\mathcal{M}_n \leq t) = \left(\sum_{k:e_n(k-1)<n} U_k(t) = 0\right)$$

for all $n \geq 1$ and $t > 0$.

Theorem 2.6.5 *Let $z \in \mathbb{R}$ be given, put*

$$t_n = \frac{1}{\theta^*}(\log(K^* n) + z),$$

and define $z_n \in [0,\theta^*)$ by $z_n = \theta^*(t_n - \lfloor t_n \rfloor)$. Then

$$\left|\left|\mathcal{D}\left(\sum_{k:e_n(k-1)<n} U_k(t_n)\right) - \mathrm{Poi}\left(\exp(-z+z_n)\right)\right|\right| \to 0.$$

In particular,

$$\mathbb{P}_i(\mathcal{M}_n \leq t_n) - \exp(-\exp(-z + z_n)) \to 0$$

for $n \to \infty$.

Remark 2.6.6 *One particularly nice feature about the Poisson approximation given in Theorem 2.6.5 is that the events we count – the number of excursions exceeding level* t_n *– can be found without knowledge of the underlying controlling Markov chain* $(J_n)_{n\geq 0}$. *Thus everything is defined completely in terms of the reflected process* $(T_n)_{n\geq 0}$.

We need the following lemma in the proof. It gives a probability estimate for the event that some arbitrary index k is contained in an excursion exceeding t.

Lemma 2.6.7 *There exist constants* K_1 *and* K_2 *such that for any* $k \leq n$ *and* $t > 0$

$$\mathbb{P}_i\left(\max_{0\leq\delta<k\leq\delta+\Delta\leq n}\sum_{r=1}^{\Delta} Z_{\delta+r} > t\right) \leq K_1(\log n)^2\exp(-\theta^* t) + K_2 n^{-2}$$

Proof: Fix any $\theta \in (0, \theta^*)$ so that $\psi(\theta) < 0$. By an exponential change of measure

$$\begin{aligned}\mathbb{P}_i\left(\sum_{r=1}^{\Delta} Z_{\delta+r} > t\right) &\leq \mathbb{E}_i^{\theta}\left(\frac{r_i^{\theta}}{r_{J_\Delta}^{\theta}}\exp(-\theta S_\Delta + \Delta\psi(\theta)); S_\Delta \geq 0\right)\\ &\leq K_2\exp(\Delta\psi(\theta)).\end{aligned}$$

With $\kappa = -4/\psi(\theta) > 0$ this implies that

$$\mathbb{P}_i\left(\max_{\substack{0\leq\delta<k\leq\delta+\Delta\leq n\\ \Delta>\kappa\log n}}\sum_{r=1}^{\Delta} Z_{\delta+r} > t\right) \leq K_2 n^{-2}.$$

When $\Delta \leq \kappa\log n$ there are at most $(\kappa\log n)^2$ possible choices of δ and Δ and using the Lundberg-type inequality (2.14) we get that

$$\mathbb{P}_i\left(\max_{\substack{0\leq\delta<k\leq\delta+\Delta\leq n\\ \Delta\leq\kappa\log n}}\sum_{r=1}^{\Delta} Z_{\delta+r} > t\right) \leq K_1(\log n)^2\exp(-\theta^* t),$$

and the result follows. □

Proof of Theorem 2.6.5: We assume that we work under the measure $\mathbb{P}_\pi$ so that the Markov chain $(J_n)_{n\geq 0}$ is stationary. One can either with a little more work extend the proof to be valid also in the non-stationary case.

Let $l = l_n \sim (\log n)^3$ be a sequence of integers, and let $I = \{0, l, 2l, \ldots, \lfloor n/l \rfloor\ l\}$ be the index set. Define the variables

$$V_k = 1\left(\max_{k<\delta\le\Delta\le k+l} \sum_{r=\delta}^{\Delta} Z_r > t_n\right)$$

for $k \in I$. We will show, using Theorem 2.6.1, that $\sum_{k\in I} V_k$ asymptotically as $n \to \infty$ has a Poisson distribution, and that it is a sufficiently good approximation of $\sum_{k:e_n(k-1)<n} U_k(t_n)$.

With $B_k = \{k-l, k, k+l\} \cap I$ for all $k \in I$, we will show that β defined by (2.34) in Theorem 2.6.1 tends to 0 for $n \to \infty$. Lemma 2.4.6 implies that $\mathbb{P}_i(V_k = 1) \le Kl\exp(-\theta^* t_n)$ for some constant K, so

$$\mathbb{E}_i(V_k)\mathbb{E}_i(V_{k'}) \le \left(\frac{K\exp(-z)}{K^*}\right)^2 l^2 n^{-2} = \tilde{K} l^2 n^{-2},$$

hence

$$\sum_{k\in I, k'\in B_k} \mathbb{E}_i(V_k)\mathbb{E}_i(V_{k'}) \le 3\tilde{K}|I|l^2 n^{-2} \sim \tilde{K} l n^{-1} \to 0$$

for $n \to \infty$. For $k' = k + l$ it follows by the Markov property that

$$\mathbb{P}_i(V_k = 1, V_{k'} = 1) = \sum_{j\in E} \mathbb{P}_i(V_k = 1, J_{k'} = j)\mathbb{P}_j(V_1 = 1).$$

So we have a similar estimate $\mathbb{E}_i(V_k V_{k'}) \le \tilde{K} l^2 n^{-2}$ as above, and the second term in β tends to 0 too for $n \to \infty$.

For the last term, note that the bivariate Markov process $(J_n, Z_n)_{n\ge1}$ under $\mathbb{P}_\pi$ can be viewed as part of doubly infinite, stationary Markov process $(J_n, Z_n)_{n\in\mathbb{Z}}$. We use the Markov property of this process to get for $A \in \mathcal{F}_{[n,\infty)}$ that

$$\begin{aligned} |\mathbb{P}(A \mid \mathcal{F}_{(-\infty,0]}) - \mathbb{P}_\pi(A)| &= |\mathbb{P}_{J_0}(A) - \mathbb{P}_\pi(A)| \\ &= \left|\sum_{j\in E}(P^n_{J_0 j} - \pi_j)\mathbb{P}_j(A)\right| \le \sum_{j\in E}|P^n_{J_0 j} - \pi_j|. \end{aligned}$$

By the exponentially fast convergence of $P^n_{ij} \to \pi_j$ for $n \to \infty$, we get exponentially fast β-mixing, i.e. for some $K, \gamma > 0$

$$\beta(n) \le K\exp(-\gamma n).$$

With $I_1 = (-\infty, k-l]$, $I_2 = [k+1, k+l]$, and $I_3 = [k+2l+1, \infty)$ we see that $\mathcal{F}_k \subseteq \mathcal{F}_{I_1\cup I_3}$ and V_k is measurable w.r.t. $\mathcal{F}_{I_2}$. From Theorem 2.6.4 we have that

$\alpha(\mathcal{F}_{I_1 \cup I_3}, \mathcal{F}_{I_2}) \leq 3\beta(l)$, and since trivially $\alpha(\mathcal{F}_k, \mathcal{F}_{I_2}) \leq \alpha(\mathcal{F}_{I_1 \cup I_3}, \mathcal{F}_{I_2})$ we obtain from Lemma 2.6.2 that

$$\mathbb{E}_\pi(|\mathbb{E}(V_k \mid \mathcal{F}_k) - \mathbb{E}_\pi(V_k)|) \leq 6\beta(l) \leq K \exp(-\gamma l) = o(n).$$

Hence the third term in β tends to zero for $n \to \infty$.

This implies that $\sum_{k \in I} V_k$ can be approximated asymptotically by a Poisson distribution. By stationarity $\mathbb{E}_\pi(V_k) = \mathbb{P}_\pi(V_1 = 1)$ and

$$\begin{aligned} \mathbb{P}_\pi(V_1 = 1) &= \mathbb{P}_\pi\left(\max_{k < \delta \leq \Delta \leq k+l} \sum_{r=\delta}^{\Delta} Z_r > t_n\right) \\ &= \mathbb{P}_\pi\left(\max_{k < \delta \leq \Delta \leq k+l} \sum_{r=\delta}^{\Delta} Z_r > t_n - \frac{z_n}{\theta^*} = \lfloor t_n \rfloor\right), \end{aligned}$$

since $Z_n \in \mathbb{Z}$ for all n. The mean of $\sum_{k \in I} V_k$ can therefore be approximated by

$$\sum_{k \in I} \mathbb{E}_\pi(V_k) \sim |I| l K^* \exp(-\theta^* t_n + z_n) \to \exp(-z + z_n),$$

which follows from Lemma 2.4.6 – using that $t_n - z_n/\theta^* = \lfloor t_n \rfloor \in \mathbb{Z}$.

To prove that $\sum_{k : e_n(k-1) < n} U_k(t_n)$ also follows a Poisson law asymptotically with mean $\exp(-z + z_n)$, we prove that

$$\mathbb{P}_i\left(\sum_{k : e_n(k-1) < n} U_k(t_n) \neq \sum_{k \in I} V_k\right) \to 0$$

and the result then follows from the coupling inequality. If (i) the start and end of any excursion exceeding t_n before time n fall within an interval of the form $\{k+1, \ldots, k+l\}$ for some $k \in I$ and (ii) no two excursions exceeding t_n start and end within the same such interval, then each excursion is clearly counted exactly once by $\sum_{k \in I} V_k$ and the two counting variables are equal. We show that the probability of (i) and (ii) tends to 1 as $n \to \infty$. The probability that some index in I is contained within an excursion exceeding level t_n is by Lemma 2.6.7 smaller than

$$|I| \left(K_1 (\log n)^2 \exp(-\theta^* t_n) + K_2 n^{-1}\right) \leq \tilde{K}(|I|(\log n)^2 n^{-1} + |I| n^{-2}) \to 0.$$

The probability that two excursions exceeding t_n both occur within an interval $\{k+1, \ldots, k+l\}$ for some $k \in I$ is smaller than

$$|I| l^2 K^2 \exp(-2\theta^* t_n) \leq \tilde{K} |I| l^2 n^{-2} \to 0.$$

Finally, the probability that an excursion exceeding t_n occurs in the last interval $\{\lfloor n/l \rfloor \; l+1, \ldots, n\}$ is smaller than $\tilde{K} l n^{-1} \to 0$. This completes the proof. □

Notes

As mentioned, one can find most of the material presented here in the literature. An excellent reference is Asmussen (2003), which contains most of the general theory about Markov additive processes. The proof of Theorem 2.4.1 is just a spelled out version of the (sketchy) proof of Theorem XIII.8.3. in Asmussen (2003). The result was first proved in Karlin & Dembo (1992) based on the asymptotic behaviour of the solution to a defective Markov renewal equation, and though their proof employs many of the same ingredients as the proof by exponential change of measure, the present proof seems more direct and slightly smoother. Studying Markov modulated queues, essentially the same result was obtained independently by Asmussen & Perry (1992). The obvious way to establish (2.24) through regenerative methods is by establishing a result like (2.20) in Theorem 2.4.3. Proposition 10.1 in Asmussen & Perry (1992) has a similar content but a different proof, and Karlin & Dembo (1992) use another approach to establish (2.24). For the ordinary random walk, Theorem 2.4.1 and (2.24) go back to Iglehart (1972). Lemma 2.4.6 is probably new in the framework of MAPs but the lemma and its proof are essentially identical to Lemma 1 by Dembo et al. (1994b), who consider only random walks. The recursive algorithm for computing the constant K^* stated in Theorem 2.5.2 was given by Karlin & Dembo (1992), although a proof that it works correctly was not included in the paper. The spectral method given by Theorem 2.5.2 is a straight forward extension of results from random walk theory going back to Wald (1947), cf. also Section VIII.5a in Asmussen (2003). In the MAP setup some complications arise as demonstrated, which make the use of the spectral method somewhat more complicated in practice. The mixing result in Theorem 2.6.4 for $\kappa = 3$ is identical to Theorem 3.1 in Takahata (1981). A general version similar to Theorem 2.6.4 was stated as Theorem 1.3.3. in Doukhan (1994). Finally, the approach taken to prove Theorem 2.6.5 is most likely new, but the result is not really new and the proof is not likely to be neither the easiest nor the best available. The approach presented was chosen solely to introduce a result and some arguments similar to those that will appear in Chapter 5 and 6.

3

Heavy Tailed Excursions

3.1 Introduction

The results obtained in Chapter 2 about the cycle maximum for a reflected MAP are all proved under the assumption that the $\Phi(\theta)$-matrix is finite. Thus the increments must have sufficiently light tails. An obvious question is how important this assumption is. If the increments are instead *heavy tailed* – where heavy tailed has a technical meaning to be defined below – we show Theorem 3.2.1 below as an analogue of Theorem 2.4.3. It turns out that in the heavy tailed case, results as well as techniques of proof come out rather differently than in the light tailed case. The consequence of Theorem 3.2.1 for the asymptotic distribution of the running maximum of the reflected MAP can, however, be derived in the same way as in the light tailed case.

To finish this chapter we discuss in Section 3.3 a different class of reflected processes – the so-called fluid models – in discrete as well as in continuous time. We give a few examples showing that the results obtained for heavy tailed Markov additive processes are useful for understanding the behaviour of fluid models with long range dependence.

3.2 Heavy tailed MAPs

Using the notation and setup from Chapter 2, we want to study the positive excursions for Markov additive processes when the H_{ij}'s are heavy tailed. The class of *subexponential distributions* is a common choice of distributions with heavy tails. A

probability measure H on $[0,\infty)$ (with tail $\overline{H}(x) = 1 - H(x)$) is subexponential if

$$\lim_{x\to\infty} \frac{\overline{H * H}(x)}{\overline{H}(x)} = \lim_{x\to\infty} \int_0^x \frac{\overline{H}(x-y)}{\overline{H}(x)} H(\mathrm{d}y) = 2.$$

A consequence of being subexponential is that $\lim_{x\to\infty} \overline{H}(x+y)/\overline{H}(x) = 1$ for all y. Also $\int \exp(\theta x) H(\mathrm{d}x) = \infty$ for all $\theta > 0$. Both properties are reasonable for what we would expect of a heavy tailed distribution. Distributions with the former property are often called *long-tailed*, and this property has the important interpretation that the conditional distribution of the overshoot of x, conditionally on getting something $> x$, tends to ∞ in law as $x \to \infty$.

We let $\mathscr{S}$ denote the set of subexponential distributions on $[0,\infty)$. For many purposes a sufficient regularity condition when studying extreme value theory is that the distribution H considered belongs to $\mathscr{S}$ and/or that the integrated tail, H_1, defined by

$$\hat{H}(x) = \int_0^x \overline{H}(y)\mathrm{d}y \quad \text{and} \quad H_1(x) = \frac{1}{\int y H(\mathrm{d}y)} \hat{H}(x),$$

belongs to $\mathscr{S}$. We will restrict our attention to another, slightly smaller class $\mathscr{S}^*$ of distributions, H, on $[0,\infty)$ having finite expectation and with the property that

$$\lim_{x\to\infty} \int_0^x \frac{\overline{H}(x-y)}{\overline{H}(x)} \overline{H}(y)\mathrm{d}y = 2\int_0^\infty \overline{H}(y)\mathrm{d}y,$$

see Klüppelberg (1988). Most notably, Theorem 3 in Klüppelberg (1988) shows that if $H \in \mathscr{S}^*$ then $H \in \mathscr{S}$ and $H_1 \in \mathscr{S}$, but we will explicitly need other properties of $\mathscr{S}^*$. Klüppelberg (1988) also identifies several classes of long-tailed distributions contained in $\mathscr{S}^*$. For instance the class of distributions with regularly varying tails.

Assume for the rest of this chapter that there exists $H \in \mathscr{S}^*$ such that with $\overline{F}_{ij}(x) = \overline{H}_{ij}(x)P_{ij}$ it holds for all $i, j \in E$ that

$$\lim_{x\to\infty} \frac{\overline{F}_{ij}(x)}{\overline{H}(x)} = \gamma_{ij} \tag{3.1}$$

for some $\gamma_{ij} \geq 0$ and at least one $\gamma_{ij} > 0$. Let Γ denote the matrix $(\gamma_{ij})_{i,j\in E}$. For future reference, it is useful to note that the convergence

$$\frac{\overline{H}(x+y)}{\overline{H}(x)} \to 1 \tag{3.2}$$

is uniform for y in a compact set (Embrechts et al. 1997, Lemma 1.3.5). Thus we get that

$$\lim_{x\to\infty} \frac{\overline{F}_{ij}(x+y)}{\overline{H}(x)} = \gamma_{ij} \tag{3.3}$$

uniformly for y in a compact set.

Recall that if $\mu < 0$ there exists a $i_0 \in E$ such that

$$\sigma = \inf\{n \geq 1 \mid T_n = 0,\ J_n = i_0\} \tag{3.4}$$

under $\mathbb{P}_{i_0}$ is an a.s. finite regeneration time for the reflected MAP $(J_n, T_n)_{n\geq 0}$, and that

$$\mathcal{M}_\sigma = \max_{0\leq n\leq \sigma} T_n \tag{3.5}$$

denotes the cycle maximum.

Theorem 3.2.1 *Under the assumption (3.1) with $H \in \mathscr{S}^*$ and $\mu < 0$*

$$\mathbb{P}_{i_0}(\mathcal{M}_\sigma > x) \sim \overline{H}(x)\mathbb{E}_{i_0}(\sigma)\,\pi\Gamma\mathbb{1} \tag{3.6}$$

for $x \to \infty$.

Remark 3.2.2 *As discussed in Remark 2.4.2, the global maximum $\mathcal{M} = \max_n S_n$ my also be of interest. In contrast to the light tailed case, the behaviour of the global maximum in the heavy tailed case is somewhat different from the behaviour of the cycle maximum. From Theorem 4 in Jelenković & Lazar (1998) we get that if $\mu < 0$, if H is long-tailed (in particular if $H \in \mathscr{S}$) and if the integrated tail $H_1 \in \mathscr{S}$, then*

$$\mathbb{P}_i(\mathcal{M} > x) \sim \frac{1}{|\mu|}\hat{H}(x)\pi\Gamma\mathbb{1} \tag{3.7}$$

for $x \to \infty$. If $H \in \mathscr{S}^*$ then as discussed H is long-tailed and $H_1 \in \mathscr{S}$ in which case (3.7) holds.

The proof of Theorem 3.2.1 is given in the next section. As argued in Section 2.4, the asymptotic distribution of the running maximum $\mathcal{M}_n = \max_{0\leq k\leq n} T_n$ follows as an easy consequence of the regenerative structure of $(J_n, T_n)_{n\geq 0}$ and Theorem 3.2.1. We give a few examples, referring to Asmussen (2003), Proposition 4.7, and Embrechts et al. (1997), Section 3.3, for further details. In particular Proposition 3.3.7 and Proposition 3.3.25 in Embrechts et al. (1997) are useful for identifying the normalisation constants.

Example 3.2.3 If H has regularly varying tail at infinity with exponent $-\alpha$, i.e.

$$\lim_{x\to\infty} \frac{\overline{H}(yx)}{\overline{H}(x)} = y^{-\alpha}$$

for all $y > 0$, then for any $i \in E$

$$\mathbb{P}_i\left(\frac{\mathcal{M}_n}{b_n} \leq x\right) \to \exp(-x^{-\alpha}), \quad n \to \infty \tag{3.8}$$

with b_n satisfying

$$n\overline{H}(b_n) \to \frac{1}{\pi\Gamma\mathbb{1}}, \quad n \to \infty.$$

One can choose $b_n = H^{\leftarrow}(1 - 1/(n\,\pi\Gamma\mathbb{1}))$ with $H^{\leftarrow}$ the generalised inverse of H defined as $H^{\leftarrow}(t) = \inf\{x \in \mathbb{R} \mid H(x) \geq t\}$. If H is tail equivalent to the power law with exponent α, i.e.

$$\lim_{x\to\infty} \overline{H}(x)x^{\alpha} = c,$$

we can choose $b_n = (nc\pi\Gamma\mathbb{1})^{1/\alpha}$. ◇

Regularly varying distributions are 'the most heavy tailed' subexponential distributions. They give rise to the Fréchet limit distribution – the extreme value distribution of type II. Other subexponential distributions with moderately heavy tails give rise to the Gumbel distribution as limiting distribution but with different normalisation compared to the light tailed case.

Example 3.2.4 If H belongs to $\mathscr{S}^*$ and in addition is tail equivalent to a Weibull distribution with parameters $\alpha > 0$ and $\beta \in (0, 1)$, i.e.

$$\lim_{x\to\infty} \overline{H}(x)\exp(\alpha x^{\beta}) = c$$

for some $c > 0$, then for any $i \in E$

$$\mathbb{P}_i\left(\frac{\mathcal{M}_n - a_n}{b_n} \leq x\right) \to \exp(-\exp(-x)), \quad n \to \infty \tag{3.9}$$

with a_n and b_n chosen as

$$a_n = \frac{1}{\alpha^{1/\beta}}(\log n)^{1/\beta}\left(1 + \beta\frac{\log(c\pi\Gamma\mathbb{1})}{\log n}\right) \quad \text{and} \quad b_n = \frac{1}{\beta\alpha^{1/\beta}}(\log n)^{1/\beta-1}.$$

Note that in the limit $\beta \to 1$, we obtain the normalisation corresponding to the light tailed case, where $\mathbb{P}_i(\mathcal{M}_\sigma > x) \sim \exp(-\theta^* x)\mathbb{E}_{i_0}(\sigma)K^*$, but the constant K^* compared to $c\pi\Gamma\mathbb{1}$ arises in quite a different way, and also the techniques of proof differ a lot. ◇

3.2.1 Proofs

The proof of Theorem 3.2.1 is divided into a number of lemmas. The idea in the proof is to use the 'one big jump' heuristic for subexponential distributions, thus an extreme value for the reflected MAP occurs due to one extreme increment. We split the extreme event $(\mathcal{M}_\sigma > x)$ into the event where the jump to a level above x happens from an intermediate level in $[x_0, x]$, $x_0 < x$, and the event where the jump happens form a level below x_0. Then we show that the probability of the first event is asymptotically negligible and that the probability of the last event has the desired asymptotic behaviour. To deal with the former we use some non-trivial downcrossing results due to Asmussen (1998) in the random walk setup – see also Foss & Zachary (2003). This argument is developed in Lemma 3.2.5 to 3.2.7. In Lemma 3.2.8 we derive the asymptotic behaviour when jumps occur from a level below some x_0 and Lemma 3.2.9 shows, using the downcrossing results, that the other probability vanishes asymptotically.

Let R be the matrix of occupation measures for the MAP, i.e. let

$$R_{ij}(D) = \sum_{n=0}^{\infty} \mathbb{P}_i(J_n = j, S_n \in D).$$

By the definition of convolution of matrix measures, the occupation measure can be written as $R = \sum_{n=0}^{\infty} F^{*n}$.

For the matrices G_- and ${}^{\#}G_+$ we define the corresponding renewal measures by

$$U_- = \sum_{n=0}^{\infty} G_-^{*n} \quad \text{and} \quad {}^{\#}U_+ = \sum_{n=0}^{\infty} \left({}^{\#}G_+\right)^{*n}.$$

And we define the occupation measure up to time τ_- as

$$R_{-,ij}(D) = \mathbb{E}_i \sum_{n=0}^{\tau_- - 1} 1(J_n = j, S_n \in D)$$

A few algebraic manipulations like those in Proposition XI.2.13 in Asmussen (2003) show that ${}^{\#}U_+ = R_-$, thus ${}^{\#}U_+$ is in particular a matrix of finite measures.

The Wiener-Hopf factorisation, Theorem 2.2.1, for MAPs implies the equation

$$R(D) = U_- * {}^{\#}U_+(D) \tag{3.10}$$

for the occupation measure, valid for $D \subseteq \mathbb{R}$ a bounded set.

The following lemma is a useful renewal theorem for MAPs.

Lemma 3.2.5 *If $\mu < 0$, it holds (in the non-lattice case) that*

$$R_{ij}\left((z, z+y]\right) \to y\frac{\pi_j}{|\mu|}, \quad z \to -\infty,$$

for all $y > 0$.

Proof: One recalls that $||G_-||$ is the transition probabilities for the Harris recurrent Markov chain $(J_{\tau_-(n)})_{n\geq 0}$. Let ν be the (unique) left invariant probability vector for $||G_-||$. The Blackwell-type renewal theorem (in the non-lattice case, cf. Example 2.1.1) gives that

$$U_{-,ij}\left((z, z+y]\right) \to y\frac{\nu_j}{|\mu_-|}, \quad z \to -\infty,$$

where $\mu_- = \sum_{i,j\in E} \nu_i \int t\, G_{-,ij}(dt)$. Observing that $U_{-,ij}$ is finite on compact intervals and 0 on $[0,\infty)$ we get that there exist constants $\alpha_1, \alpha_2 > 0$ such that for all $i, j \in E$, $y > 0$ and $z \in \mathbb{R}$

$$U_{-,ij}\left((z, z+y]\right) \leq \alpha_1 + \alpha_2 y. \tag{3.11}$$

The Wiener-Hopf identity (3.10) together with dominated convergence imply that

$$\begin{aligned} R_{ij}\left((z, z+y]\right) &= \sum_{k\in E} U_{-,ik} * {}^{\#}U_{+,kj}\left((z, z+y]\right) \\ &= \sum_{k\in E} {}^{\#}U_{+,kj} * U_{-,ik}\left((z, z+y]\right) \\ &\to \frac{y}{|\mu_-|}\sum_{k\in E} \nu_k ||{}^{\#}U_+||_{kj}, \quad z \to -\infty. \end{aligned} \tag{3.12}$$

From the Wiener-Hopf identity

$$I - P = I - ||F|| = (I - ||{}^{\#}G_+||)(I - ||G_-||)$$

it follows by multiplying with π from the left that $\nu \propto \pi(I - ||{}^{\#}G_+||) \neq 0$ or

$$\pi \propto \nu||{}^{\#}U_+||.$$

Using that ${}^{\#}U_+ = R_-$ where $\sum_{j\in E} ||R_{-,ij}|| = \mathbb{E}_i(\tau_-)$, the constant of proportionality is seen to be $\sum_{i\in E} \nu_i \mathbb{E}_i(\tau_-) = \mathbb{E}_\nu(\tau_-)$, hence $\sum_{k\in E} \nu_k ||{}^{\#}U_+||_{kj} = \mathbb{E}_\nu(\tau_-)\pi_j$. The result now follows from the Wald identity $\mathbb{E}_\nu(\tau_-)\mu = \mu_-$. □

Of course a similar result holds in the lattice case. We skip the details.

For $x > 0$ let

$$N_\sigma(x) = \sum_{n=0}^{\sigma-1} 1(T_n > x, T_{n+1} \leq x)$$

be the number of downcrossings of level x before time σ. Let

$$\rho(x) = \inf\{n \geq 1 \mid S_n \leq -x\}$$

be the time of the first downcrossing for the MAP of level $-x$, and if also $y > 0$ denote by

$$N(x,y) = \sum_{n=0}^{\rho(x+y)-1} 1(S_n > -x, S_{n+1} \leq -x)$$

the number of downcrossings of level $-x$ *before* a level below $-(x+y)$ is reached. Note that

$$N(x,y) \nearrow N(x) = \sum_{n=0}^{\infty} 1(S_n > -x, S_{n+1} \leq -x)$$

for $y \to \infty$, where $N(x)$ is the *total* number of downcrossings of level $-x$. Let in the following m_- be the matrix $(m_{-,ij})_{i,j\in E}$ given by

$$m_{-,ij} = \int_0^\infty F_{ij}(-z)\mathrm{d}z.$$

Lemma 3.2.6 *When $\mu < 0$ it holds for all $i \in E$ that*

$$\mathbb{E}_i N(x) \to \frac{\pi m_- \mathbb{1}}{|\mu|}, \quad x \to \infty,$$

and the convergence

$$\mathbb{E}_i N(x,y) \nearrow \mathbb{E}_i N(x), \quad y \to \infty$$

is uniform in x.

Proof: By conditioning on the value of (J_n, S_n) for $n \geq 0$ we get using Lemma 3.2.5 that for all $i \in E$

$$\begin{aligned}
\mathbb{E}_i N(x) &= \sum_{j,k\in E} \int_{-x}^\infty F_{jk}(-(x+y)) R_{ij}(\mathrm{d}y) \\
&= \sum_{j,k\in E} \int_0^\infty F_{jk}(-z) R_{ij}(\mathrm{d}z - x) \qquad (3.13) \\
&\to \sum_{j,k\in E} \frac{\pi_j}{|\mu|} \int_0^\infty F_{jk}(-z)\mathrm{d}z, \quad x \to \infty \\
&= \frac{\pi m_- \mathbb{1}}{|\mu|}.
\end{aligned}$$

To prove the uniform convergence, use (3.11) combined with (3.12) to find an α such that $R_{jk}([z, z+1)) \leq \alpha$ for all z, then using (3.13) above,

$$\begin{aligned}\mathbb{E}_i N(x) &\leq \sum_{j,k\in E}\sum_{n=0}^{\infty} F_{jk}(-n)R_{ij}([n-x, n-x+1)) \\ &\leq \alpha \sum_{j,k\in E}\sum_{n=0}^{\infty} F_{jk}(-n) < \infty.\end{aligned}$$

The right hand side is finite due to H_{ij} having finite mean and it is independent of x. With $\mathcal{M} = \max_{n\geq 0} S_n$ and using the strong Markov property we obtain

$$\begin{aligned}0 &\leq \mathbb{E}_i N(x) - \mathbb{E}_i N(x,y) \\ &= \mathbb{E}_i \sum_{n=\rho(x+y)}^{\infty} 1_{(S_n > -x, S_{n+1}\leq -x)} \\ &\leq \max_j \mathbb{P}_j(\mathcal{M} > y) \sup_z \mathbb{E}_j(N(z)) \to 0\end{aligned}$$

uniformly in x since $\mathcal{M} < \infty$ almost surely. □

Lemma 3.2.7 *Under the assumption (3.1) and $\mu < 0$*

$$\lim_{x\to\infty} \frac{\mathbb{E}_{i_0} N_\sigma(x)}{\overline{H}(x)} = \frac{\mathbb{E}_{i_0}(\sigma)}{|\mu|}\, \pi\Gamma\mathbb{1}\, \pi m_- \mathbb{1}.$$

Proof: By the regeneration property of $(J_n, T_n)_{n\geq 0}$ the invariant distribution, λ, for the reflected MAP can be represented as

$$\lambda(i, A) = \frac{1}{\mathbb{E}_{i_0}(\sigma)} \mathbb{E}_{i_0}\left(\sum_{k=0}^{\sigma-1} 1(J_k = i, T_k \in A)\right),$$

for $A \in \mathbb{B}$, which gives that

$$\mathbb{E}_{i_0} N_\sigma(x) = \mathbb{E}_{i_0}(\sigma) \sum_{i,j\in E} \int_x^\infty F_{ij}(x-y)\lambda(i, \mathrm{d}y). \tag{3.14}$$

By Remark 2.4.2, cf. also Remark 3.2.2, the invariant distribution λ coincides with the distribution of $(\overleftarrow{J_0}, \overleftarrow{\mathcal{M}})$ if $\overleftarrow{J_0}$ has distribution π. Since

$$\lim_{x\to\infty} \frac{\overleftarrow{P}_{ij}(1 - \overleftarrow{H}_{ij}(x))}{\overline{H}(x)} = \lim_{x\to\infty} \frac{\pi_j \overline{F}_{ji}(x)/\pi_i}{\overline{H}(x)} = \frac{\pi_j \gamma_{ji}}{\pi_i}.$$

we obtain from (3.7) in Remark 3.2.2 that

$$\frac{\lambda(i,(x,\infty))}{\hat{H}(x)} = \frac{\mathbb{P}_\pi(\overleftarrow{J_0}=i,\overleftarrow{\mathcal{M}}>x)}{\hat{H}(x)} \to \frac{\pi_i}{|\mu|}\pi\Gamma\mathbb{1}, \quad x\to\infty.$$

The right tail of $\lambda(i,\cdot)$ is thus asymptotically equivalent to a measure proportional to $\overline{H}(x)\mathrm{d}x$, cf. the definition of $\hat{H}(x)$. Since $F_{ij}(x)=\int_{-\infty}^x F_{ij}(\mathrm{d}z)$, we get by interchanging the order of integration that

$$\begin{aligned}\int_x^\infty F_{ij}(x-y)\overline{H}(y)\mathrm{d}y &= \int_{-\infty}^0\int_x^{x-z}\overline{H}(y)\mathrm{d}yF_{ij}(\mathrm{d}z)\\ &\sim \overline{H}(x)\int_{-\infty}^0 -zF_{ij}(\mathrm{d}z)=\overline{H}(x)\int_0^{-\infty}F_{ij}(-z)\mathrm{d}z\end{aligned}$$

for $x\to\infty$. Here we use that $\int_x^{x-z}\overline{H}(y)\mathrm{d}y\sim -z\overline{H}(x)$ by (3.2). Thus if we can just substitute $\lambda(i,\cdot)$ in the limit with the tail equivalent measure, we get that for all $i,j\in E$.

$$\frac{\int_x^\infty F_{ij}(x-y)\lambda(i,\mathrm{d}y)}{\overline{H}(x)} \to \frac{\pi\Gamma\mathbb{1}}{|\mu|}\pi_i\int_0^\infty F_{ij}(-z)\mathrm{d}z, \quad x\to\infty. \tag{3.15}$$

To formalise the argument, Corollary 1 in Asmussen et al. (2002) implies that the substitution of $\lambda(i,\cdot)$ is indeed valid – provided as assumed that $H\in\mathscr{S}^*$.

Using (3.14), we conclude by summing over i,j in (3.15) that

$$\frac{\mathbb{E}_{i_0}N_\sigma(x)}{\overline{H}(x)} \to \frac{\mathbb{E}_{i_0}(\sigma)}{|\mu|}\,\pi\Gamma\mathbb{1}\,\pi m_-\mathbb{1}, \quad x\to\infty.$$

□

Let

$$\tau(x)=\inf\{n\ge 1 \mid T_n>x\}$$

so that $(\mathcal{M}_\sigma>x)=(\tau(x)<\sigma)$, and let for $x_0<x$, $y_0\ge 0$

$$A(x,x_0,y_0) = (\tau(x)<\sigma, T_{\tau(x)}>x+y_0, T_{\tau(x)-1}<x_0).$$

That is, $A(x,x_0,y_0)$ is the event that the T-process will exceed x before time σ, and when doing so the process jumps from a value below x_0 to a value above $x+y_0$.

Lemma 3.2.8 *Under the assumption (3.1) and $\mu<0$,*

$$\lim_{x_0\to\infty}\lim_{x\to\infty}\frac{\mathbb{P}_{i_0}(A(x,x_0,y_0))}{\overline{H}(x)} = \mathbb{E}_{i_0}(\sigma)\,\pi\Gamma\mathbb{1} \tag{3.16}$$

for all $y_0\ge 0$.

Proof: Put $\sigma(x) = \sigma \wedge \tau(x)$. Then

$$\begin{aligned}\mathbb{P}_{i_0}(A(x,x_0,y_0)) &= \sum_{n=1}^{\infty} \mathbb{P}_{i_0}(T_n > x+y_0, T_{n-1} < x_0, \sigma(x) \geq n) \\ &= \sum_{i,j\in E}\sum_{n=1}^{\infty} \mathbb{P}_{i_0}(T_n > x+y_0, T_{n-1} < x_0, J_{n-1}=i, J_n = j, \sigma(x) \geq n) \\ &= \sum_{i,j\in E}\sum_{n=1}^{\infty} \int_0^{x_0} \overline{F}_{ij}(x+y_0-y)\mathbb{P}_{i_0}(T_{n-1} \in \mathrm{d}y, J_{n-1}=i, \sigma(x)\geq n).\end{aligned}$$

Using that the convergence in (3.3) is uniform for $y \in [0, x_0]$ and that $\sigma(x) \nearrow \sigma$ for $x \to \infty$ we get that

$$\frac{\mathbb{P}_{i_0}(A(x,x_0,y_0))}{\overline{H}(x)} \to \sum_{i,j\in E}\sum_{n=1}^{\infty} \gamma_{ij}\mathbb{P}_{i_0}(T_{n-1} < x_0, J_{n-1}=i, \sigma \geq n).$$

Letting $x_0 \to \infty$, the result follows by using that

$$\sum_{n=1}^{\infty} \mathbb{P}_{i_0}(J_{n-1}=i, \sigma \geq n) = \sum_{n=0}^{\sigma-1} \mathbb{P}_{i_0}(J_n = i) = \pi_i \mathbb{E}_{i_0}(\sigma).$$

□

For $x_0 < x$ let $A(x,x_0) = A(x,x_0,0) = (\tau(x) < \sigma, T_{\tau(x)-1} < x_0)$ and let $B(x,x_0) = (\tau(x) < \sigma, T_{\tau(x)-1} \geq x_0)$.

Lemma 3.2.9 *Under the assumption (3.1) and $\mu < 0$,*

$$\lim_{x_0\to\infty} \limsup_{x\to\infty} \frac{\mathbb{P}_{i_0}(B(x,x_0))}{\overline{H}(x)} = 0. \tag{3.17}$$

Proof: With $\kappa = \pi m_- \mathbb{1}/|\mu|$ and for $\varepsilon > 0$ be given, choose x and y_0 large enough such that according to Lemma 3.2.6

$$\mathbb{E}_j(N(y-x,x)) \geq \kappa - \varepsilon$$

for $y \geq x+y_0$ and all $j \in E$. On the event $(\tau(x) < \sigma)$, the number of times T_n crosses level x from above after time $\tau(x)$ and before time σ is larger than the number of times T_n crosses level x from above after time $\tau(x)$ and before hitting zero. Hence

the strong Markov property of $(J_n, T_n)_{n\geq 0}$ gives

$$\begin{aligned}
&\mathbb{E}_{i_0}(N_\sigma(x); A(x,x_0)) \;\geq\; \mathbb{E}_{i_0}(\mathbb{E}_{J_{\tau(x)}}(N(T_{\tau(x)}-x,x)); A(x,x_0)) \\
&= \sum_{j\in E}\int_x^\infty \mathbb{E}_j(N(y-x,x))\mathbb{P}_{i_0}(T_{\tau(x)}\in \mathrm{d}y, J_{\tau(x)}=j, A(x,x_0)) \\
&\geq \sum_{j\in E}\int_{x+y_0}^\infty \mathbb{E}_j(N(y-x,x))\mathbb{P}_{i_0}(T_{\tau(x)}\in \mathrm{d}y, J_{\tau(x)}=j, A(x,x_0)) \\
&\geq (\kappa-\varepsilon)\mathbb{P}_{i_0}(A(x,x_0,y_0)).
\end{aligned}$$

Using Lemma 3.2.8

$$\lim_{x_0\to\infty}\liminf_{x\to\infty}\frac{\mathbb{E}_{i_0}(N_\sigma(x); A(x,x_0))}{\overline{H}(x)} \geq \mathbb{E}_{i_0}(\sigma)(\kappa-\varepsilon)\,\pi\Gamma\mathbb{1}$$

Since $\varepsilon > 0$ was arbitrary and

$$1(B(x,x_0)) \leq N_\sigma(x)1(B(x,x_0)) = N_\sigma(x) - N_\sigma(x)1(A(x,x_0))$$

we get from Lemma 3.2.7 that

$$\lim_{x_0\to\infty}\limsup_{x\to\infty}\frac{\mathbb{P}_{i_0}(B(x,x_0))}{\overline{H}(x)} \;\leq\; \mathbb{E}_{i_0}(\sigma)\,\pi\Gamma\mathbb{1}\left(\frac{\pi m_-\mathbb{1}}{|\mu|}-\kappa\right) = 0.$$

□

Proof of Theorem 3.2.1 From the identity

$$\mathbb{P}_{i_0}(\mathcal{M}_\sigma > x) = \mathbb{P}_{i_0}(A(x,x_0)) + \mathbb{P}_{i_0}(B(x,x_0)).$$

the result (3.6) follows immediately from Lemma 3.2.8 and 3.2.9. □

3.3 Fluid models

We will consider a different class of models where the same kind of heavy tailed behaviour can occur, but instead of occurring directly as a consequence of one big jump the heavy tails occur due to an aggregation of (smaller) increments over a very long periods. We will consider discrete as well as continuous time processes, and we choose not to distinguish between them in notation. Thus $t \geq 0$ means either $t \in \mathbb{N}_0$ or $t \in [0,\infty)$. In continuous time we always assume that the processes are cadlag. For a real-valued process $(S_t)_{t\geq 0}$ with $S_0 = 0$ we define its reflection, $(T_t)_{t\geq 0}$, at the zero barrier by

$$T_t = S_t - \inf_{0\leq s\leq t} S_s. \tag{3.18}$$

This definition coincides by (2.10) with the definition of the reflection of a discrete time MAP. To relate general continuous time processes to the results of the previous section, we need the following definition.

Definition 3.3.1 *An increasing sequence of stopping times $(\rho_n)_{n\geq 0}$ with $\rho_0 = 0$ is called a sequence of Markov additive sampling times if there exists a Markov chain $(J_n)_{n\geq 0}$ with finite state space E such that the process $(S_{\rho_n}, J_n)_{n\geq 0}$ is a MAP.*

By the definition (3.18), the reflection of a sampled process is always less than sampling the reflection of the original process. That is, if $(T_n^\rho)_{n\geq 0}$ denotes the reflection of $(S_{\rho_n})_{n\geq 0}$ for an increasing sequence of stopping times $\rho = (\rho_n)_{n\geq 0}$, it holds for all ρ that

$$T_{\rho_n} \geq T_n^\rho \quad \text{for all} \quad n \geq 0. \tag{3.19}$$

Thus if there is a sequence of Markov additive sampling times, we immediately get a lower bound on the cycle maximum of the original reflected process in terms of the cycle maximum for the reflected Markov additive process. We claim that in some cases this bound is sharp, and we give examples for which the bound is always an equality. For these examples the most difficult problem is to verify that relevant properties of the distributions of the increments $S_{\rho_n} - S_{\rho_{n-1}}$ hold for the sampled MAP.

A suitable framework for examples is that of *fluid models* controlled by an underlying process $(J_t)_{t\geq 0}$, which we call the state process, taking values in a finite set E. We will be able to find a sequence of Markov sampling times for such fluid models if the state process is a semi-Markov chain and if the fluid rates (to be defined below) are conditionally independent given the state process. Defining the *jump times*, $(\rho_n)_{n\geq 0}$, for the state process by $\rho_0 = 0$ and

$$\rho_n = \inf\{t > \rho_{n-1} \mid J_t \neq J_{\rho_{n-1}}\}, \ n \geq 1,$$

the state process $(J_t)_{t\geq 0}$ is a semi-Markov chain if $(J_{\rho_n}, \rho_n)_{n\geq 0}$ is a Markov renewal process. In other words, if $(\widetilde{J}_n, \rho_n)_{n\geq 0}$ denotes a Markov renewal process, there is a corresponding semi-Markov chain $(J_t)_{t\geq 0}$ defined by

$$J_t = \widetilde{J}_n \quad \text{for} \quad \rho_n \leq t < \rho_{n+1}.$$

We usually call $\rho_{n+1} - \rho_n$ the *duration* that the semi-Markov chain stays in the state J_{ρ_n}, and the conditional distributions of $\rho_{n+1} - \rho_n$ given J_{ρ_n} and $J_{\rho_{n+1}}$ are called the *duration distributions*. We assume in the following that $(J_t)_{t\geq 0}$ is a semi-Markov chain such that the sampled Markov chain $(J_{\rho_n})_{n\geq 0}$ has transition probabilities given by P, which is irreducible and aperiodic with invariant probability measure π. The *fluid rates* are assumed to be given by a sequence of processes $\widetilde{\mathbb{X}}_n = (\widetilde{X}_s^n)_{0\leq s<\rho_{n+1}-\rho_n}$

for $n = 0, 1, \ldots$, which are conditionally independent given $(J_{\rho_n})_{n\geq 0}$ and such that the conditional distribution of $\widetilde{\mathbb{X}}_n$ depends on J_{ρ_n} only. Thus the fluid rates consist of a sequence of stochastic real valued functions defined on the stochastic intervals $[0, \rho_{n+1} - \rho_n)$. The fluid rate process $(X_t)_{t\geq 0}$ is then defined as

$$X_t = \widetilde{X}^n_{t-\rho_n} \quad \text{if} \quad \rho_n \leq t < \rho_{n+1}$$

and we define $(S_t)_{t\geq 0}$ by

$$S_t = \int_0^t X_t dt \tag{3.20}$$

with summation substituting integration in discrete time. The reflection, $(T_t)_{t\geq 0}$, of $(S_t)_{t\geq 0}$ at the zero-barrier is then called the fluid model controlled by $(J_t)_{t\geq 0}$ with fluid rate $(X_t)_{t\geq 0}$. The interpretation is that of a container where the fluid flow is given by $(X_t)_{t\geq 0}$, which is controlled by the underlying process $(J_t)_{t\geq 0}$. Whenever the container becomes empty, nothing flows out and it remains empty until something flows in again, hence the reflection at zero.

Due to irreducibility there is an i_0 such that

$$\sigma = \inf\{n \geq 1 \mid J_{\rho_n} = i_0, T^\rho_n = 0\}$$

is a regeneration time for the reflection of the sampled MAP $(J_{\rho_n}, T^\rho_n)_{n\geq 0}$. Defining

$$\widetilde{\sigma} = \inf\{t > 0 \mid J_t = i_0, J_{t-} \neq i_0, T_t = 0\}$$

we have $\widetilde{\sigma} \geq \rho_\sigma$ and $\widetilde{\sigma}$ is a regeneration time for $(J_t, T_t)_{t\geq 0}$ if $\widetilde{\sigma} < \infty$ a.s. Introduce also

$$\mathcal{M}^\rho_\sigma := \max_{0\leq n\leq \sigma} T^\rho_n$$

and if $\widetilde{\sigma} < \infty$ a.s.

$$\mathcal{M}_{\widetilde{\sigma}} := \max_{0\leq t\leq \widetilde{\sigma}} T_t,$$

for which we in general have $\mathcal{M}_{\widetilde{\sigma}} \geq \mathcal{M}^\rho_\sigma$.

Lemma 3.3.2 *If the fluid rate process has constant sign in between jumps, i.e. if for all $n \geq 0$ either*

$$\begin{aligned} \widetilde{X}^n_s &\geq 0 \quad \textit{for all} \quad s \in [0, \rho_{n+1} - \rho_n) \quad \textit{or} \\ \widetilde{X}^n_s &\leq 0 \quad \textit{for all} \quad s \in [0, \rho_{n+1} - \rho_n) \end{aligned}$$

we have that

$$\widetilde{\sigma} = \rho_\sigma \quad \textit{and} \quad \mathcal{M}_{\widetilde{\sigma}} = \mathcal{M}^\rho_\sigma. \tag{3.21}$$

Proof: By the constant sign condition imposed, the process $(S_t)_{t\geq 0}$ (as well as its reflection) is monotone in between the jumps of the underlying semi-Markov process, hence the infimum up to a jump time ρ_n must occur at another jump time ρ_k, $k \in \{0, \ldots, \rho_n\}$. This implies that $T^\rho_n = T_{\rho_n}$ for all $n \geq 0$. Therefore, since $\widetilde{\sigma}$ necessarily is a jump time, it equals ρ_σ. Finally, again by monotonicity, the reflected process takes its maximum $\mathcal{M}_{\widetilde{\sigma}}$ over the time interval $[0, \widetilde{\sigma}]$ at a jump time, hence the maximum equals $\mathcal{M}^\rho_\sigma$. □

This lemma gives a class of fluid models for which the asymptotic behaviour of the cycle maximum can be treated very easily by e.g. Theorem 3.2.1, if we can verify the necessary conditions for the increment distributions for the sampled MAP. In general the increment distributions, $(H_{ij})_{i,j\in E}$, for the sampled MAP are given by

$$H_{ij}(x) = \mathbb{P}_i\left(\int_0^{\rho_1} X_s \mathrm{d}s \leq x \mid J_1 = j\right).$$

Example 3.3.3 The simplest case is when the rates are constant in between jumps. That is, assume that $\widetilde{X}^n_s = \widetilde{X}^n_0$ for all $s \in [0, \rho_{n+1} - \rho_n)$. Assume that the conditional distribution of $\widetilde{X}^n_0$ given $J_n = i$ is K_i with mean ζ_i, and that the conditional distribution of $\rho_{n+1} - \rho_n$ given $J_n = i$ and $J_{n+1} = j$ is L_{ij} with mean ξ_{ij}. The increment distribution H_{ij} then has mean

$$\mu_{ij} = \zeta_i \xi_{ij},$$

for $i, j \in E$. Therefore the invariant mean drift for the sampled MAP is

$$\mu = \sum_{i,j} \pi_i P_{ij} \zeta_i \xi_{ij}.$$

We assume that $\mu < 0$ and assume, furthermore, that there exists a distribution function L with *regularly varying* tail with index $-\alpha$ such that

$$\frac{\overline{L}_{ij}(x)}{\overline{L}(x)} \to \delta_{ij},$$

with $\delta_{ij} \in [0, \infty)$ and at least one $\delta_{ij} > 0$. Clearly

$$\overline{H}_{ij}(x) = \int_0^\infty \overline{L}_{ij}\left(\frac{x}{y}\right) K_i(\mathrm{d}y),$$

hence if the limit $\overline{L}_{ij}(x/y)/\overline{L}(x) \to \delta_{ij} y^\alpha$ for $x \to \infty$ has an integrable majorant (or the limit and integration can be interchanged by other means), we obtain that

$$\frac{\overline{H}_{ij}(x)}{\overline{L}(x)} \to \delta_{ij} \int_0^\infty y^\alpha K_i(\mathrm{d}y).$$

With $\zeta_{i,\alpha} = \int_0^\infty y^\alpha K_i(\mathrm{d}y) < \infty$, Theorem 3.2.1 holds for the reflection of the sampled MAP with $H = L$ and $\gamma_{ij} = \delta_{ij}\zeta_{i,\alpha}$, hence by the equalities $\widetilde{\sigma} = \rho_\sigma$ and $\mathcal{M}_{\widetilde{\sigma}} = \mathcal{M}^\rho_\sigma$ we obtain

$$\lim_{x\to\infty} \frac{\mathbb{P}_{i_0}(\mathcal{M}_{\widetilde{\sigma}} > x)}{\overline{L}(x)} = \mathbb{E}_{i_0}(\widetilde{\sigma}) \frac{\sum_{i,j} \pi_i \delta_{ij} \zeta_{i,\alpha}}{\sum_i \pi_i P_{ij} \xi_{ij}}.$$

Here we used Walds identity for MAPs to identify $\mathbb{E}_{i_0}(\widetilde{\sigma}) = \mathbb{E}_{i_0}(\sigma) \sum_{i\in E} \pi_i P_{ij} \xi_{ij}$. $\diamond$

One can generalise in many directions. We will give one generalisation which allows for a finite (stochastic) number of changes in the flow rate in between jumps. To deal with this slightly more complicated situation we make a digression to a general result on the summation of a stochastic heavy tailed number of positive stochastic variables.

Suppose that K is a distribution on $(0,\infty)$ and G a distribution on $\mathbb{N}$. Define

$$H(A) = \sum_{k=1}^{\infty} K^{*n}(A) G(n).$$

If ρ is a random variable with distribution G independent of a sequence of *iid* variables $(Z_n)_{n\geq 1}$ each with distribution K, then

$$H(A) = \mathbb{P}\left(\sum_{k=1}^{\rho} Z_n \in A \right).$$

Let $(N_x)_{x\geq 0}$ denote the counting process corresponding to the renewal process given by $(Z_n)_{n\geq 1}$. That is, with $S_n = \sum_{k=1}^n Z_k$ we define $(N_x)_{x\geq 0}$ by

$$N_x = \sum_{n=1}^{\infty} 1_{(S_n \leq x)},$$

so that $N_x \leq n$ if and only if $S_{n+1} > x$. We observe that the tail of H can be written as

$$\overline{H}(x) = \mathbb{P}\left(\sum_{k=1}^{\rho} Z_n > x \right) = \mathbb{P}(N_x < \rho) = \mathbb{E}(\overline{G}(N_x)).$$

Lemma 3.3.4 *If $\overline{G}$ is regularly varying at infinity with exponent $-\alpha$, K has mean value ζ, and for all $\varepsilon > 0$*

$$\mathbb{P}\left(\left| \frac{N_x}{x} - \frac{1}{\zeta} \right| > \varepsilon \right) = o(\overline{G}(x)) \tag{3.22}$$

for $x \to \infty$, then

$$\overline{H}(x) \sim \zeta^\alpha \overline{G}(x)$$

and $\overline{H}$ is regularly varying at infinity with exponent $-\alpha$.

Proof: For $\varepsilon > 0$ given we decompose

$$\mathbb{E}(\overline{G}(N_x)) = \mathbb{E}\left(\overline{G}(N_x); \left|\frac{N_x}{x} - \frac{1}{\zeta}\right| \le \varepsilon\right) + \mathbb{E}\left(\overline{G}(N_x); \left|\frac{N_x}{x} - \frac{1}{\zeta}\right| > \varepsilon\right).$$

Using that $\overline{G}(N_x) \le 1$, the last term divided by $\overline{G}(x)$ tends to zero for $x \to \infty$ by (3.22). For the other term observe that $\overline{G}(x(1/\zeta + \varepsilon)) \le \overline{G}(N_x) \le \overline{G}(x(1/\zeta - \varepsilon))$ on $\left|\frac{N_x}{x} - \frac{1}{\zeta}\right| \le \varepsilon$, hence by regular variation of $\overline{G}$

$$\left(\frac{1}{\zeta} + \varepsilon\right)^{-\alpha} \le \liminf_{x\to\infty} \frac{\mathbb{E}(\overline{G}(N_x))}{\overline{G}(x)} \le \limsup_{x\to\infty} \frac{\mathbb{E}(\overline{G}(N_x))}{\overline{G}(x)} \le \left(\frac{1}{\zeta} - \varepsilon\right)^{-\alpha}.$$

Letting $\varepsilon \to 0$ the result follows. □

A useful sufficient criteria for condition (3.22) is the following moment condition.

Lemma 3.3.5 *Condition (3.22) of the previous lemma is fulfilled if*

$$\int x^{2\alpha+\delta} K(\mathrm{d}x) < \infty \tag{3.23}$$

for some $\delta > 0$ with $2\alpha + \delta \ge 2$.

Proof: The condition (3.22) is a restriction on the *large deviations* of N_x from $x\zeta$. Letting $n_x = \lceil x(1/\zeta + \varepsilon)\rceil$ and $m_x = \lceil x(1/\zeta - \varepsilon)\rceil$, we see that

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{N_x}{x} - \frac{1}{\zeta}\right| > \varepsilon\right) &= \mathbb{P}\left(N_x > x\left(\frac{1}{\zeta} + \varepsilon\right)\right) + \mathbb{P}\left(N_x < x\left(\frac{1}{\zeta} - \varepsilon\right)\right) \\
&\le \mathbb{P}\left(S_{n_x} \le x\right) + \mathbb{P}\left(S_{m_x} \ge x\right) \\
&\le \mathbb{P}\left(|S_{n_x} - n_x\zeta| \ge x\varepsilon\zeta\right) + \mathbb{P}\left(|S_{m_x} - m_x\zeta| \ge (x\varepsilon - 1)\zeta\right),
\end{aligned}$$

Assuming p'th moments for some $p \ge 1$ and using Markov's inequality, we find that

$$\mathbb{P}\left(|S_{n_x} - n_x\zeta| \ge x\varepsilon\zeta\right) \le \frac{\mathbb{E}(|S_{n_x} - n_x\zeta|^p)}{(x\varepsilon\zeta)^p}.$$

With $X_1, \ldots, X_n$ *iid* variables with mean 0 and p'th moment for $p \ge 2$ the inequality

$$\mathbb{E}\left(\left|\sum_{k=1}^{n} X_k\right|^p\right) \le C(p) n^{p/2} \mathbb{E}|X_1|^p$$

holds for some constant $C(p)$ depending only on p (Petrov 1995, Theorem 2.10). Hence

$$\begin{aligned}
\mathbb{P}\left(|S_{n_x} - n_x\zeta| \ge x\varepsilon\zeta\right) &\le \frac{C(p) n_x^{p/2} \mathbb{E}|Z_1 - \zeta|^p}{(x\varepsilon\zeta)^p} \\
&\le C(p,\varepsilon) x^{-p/2},
\end{aligned}$$

valid for x large enough and some constant $C(p,\varepsilon)$ not depending on x. A similar estimate holds for $\mathbb{P}\left(|S_{m_x} - m_x\zeta| \geq (x\varepsilon - 1)\zeta\right)$. Therefore, if

$$\mathbb{E}Z_1^{2\alpha+\delta} = \int x^{2\alpha+\delta} K(\mathrm{d}x) < \infty$$

for some $\delta > 0$ and $2\alpha + \delta \geq 2$,

$$\mathbb{P}\left(\left|\frac{N_x}{x} - \frac{1}{\zeta}\right| > \varepsilon\right) \leq O(x^{-\alpha-\delta/2}) = o(\overline{G}(x))$$

if $\overline{G}(x)$ is regularly varying at infinity with exponent $-\alpha$. □

Example 3.3.6 Suppose that $(G_{ij})_{i,j\in E}$ are distributions on $\mathbb{N}$ with mean ξ_{ij}, that $(R_i)_{i\in E}$ are distributions on $[0,\infty)$ with mean λ_i and that $(K_i)_{i\in E}$ are distributions on $\mathbb{R}$ with mean ζ_i. Given $J_n = i$ and $J_{n+1} = j$ we define the conditional distribution of $\rho_{n+1} - \rho_n$ and $\mathbb{X}_n$ as follows. Let η have distribution G_{ij}, let $(\kappa_k)_{k=1}^{\eta}$ be a sequence of *iid* stochastic variables (independent of η) with distribution R_i and let $(Z_k)_{k=1}^{\eta}$ be a sequence of *iid* stochastic variables (independent of η) with distribution K_i. Then let

$$\rho_{n+1} = \sum_{k=1}^{\eta} \kappa_k + \rho_n$$

and define $\widetilde{\mathbb{X}}_n$ by

$$\widetilde{X}_n^s = Z_k \quad \text{if} \quad \sum_{l=1}^{k-1} \kappa_n \leq s < \sum_{l=1}^{k} \kappa_n.$$

Thus the duration distributions $(L_{ij})_{i,j\in E}$ are given by

$$L_{ij}(A) = \sum_{k=1}^{\infty} R_i^{*n}(A) G_{ij}(n),$$

and the increment distributions $(H_{ij})_{i,j\in E}$ are given by

$$H_{ij}(A) = \sum_{k=1}^{\infty} \widetilde{K}_i^{*n}(A) G_{ij}(n), \quad \widetilde{K}_i(x) = \int K_i\left(\frac{x}{y}\right) R_i(\mathrm{d}y).$$

The interpretation is as follows. In between jumps, a conditionally independent renewal process given by the κ's marks the times at which the fluid rate changes, each time the change is drawn independently from K_i. The stochastic variable η is the

number of renewals encountered before a jump, and jumps take place immediately after renewal number η. We observe that

$$\mu_{ij} = \zeta_i \lambda_i \xi_{ij}$$

so the invariant mean for the sampled MAP is

$$\mu = \sum_{i,j} \pi_i P_{ij} \zeta_i \lambda_i \xi_{ij},$$

which we assume to be < 0. To stay within the framework of Lemma 3.3.2 we need some further assumptions on the distributions K_i for $i \in E$. We will assume that $E = E^+ \cup E^-$ is a partition of the state space E into two disjoint subsets such that

$$\begin{aligned} K_i((0,\infty)) &= 1 \quad \text{for} \quad i \in E^+ \\ K_i((-\infty,0]) &= 1 \quad \text{for} \quad i \in E^-. \end{aligned}$$

In this case the fluid rates have constant sign in between jumps and Lemma 3.3.2 holds. Assume that there is distribution function G with regularly varying tail with exponent $-\alpha$ such that

$$\lim_{x\to\infty} \frac{\overline{G}_{ij}(x)}{\overline{G}(x)} = \delta_{ij} \tag{3.24}$$

for all $i, j \in E$, $\delta_{ij} \in [0,\infty)$ and at least one $\delta_{ij} > 0$ for $i \in E^+$. Assume also that the moment condition (3.23) is satisfied for K_i and R_i for all $i \in E^+$, in which case it holds for $\widetilde{K}_i$ for $i \in E^+$ also. Hence by Lemma 3.3.4

$$\lim_{x\to\infty} \frac{\overline{H}_{ij}(x)}{\overline{G}(x)} = \delta_{ij} \zeta_i^\alpha \lambda_i^\alpha,$$

so Theorem 3.2.1 holds with $H = G$ and $\gamma_{ij} = \delta_{ij}\zeta_i^\alpha \lambda_i^\alpha$. We get that

$$\lim_{x\to\infty} \frac{\mathbb{P}_{i_0}(\mathcal{M}_{\widetilde{\sigma}} > x)}{\overline{G}(x)} = \mathbb{E}_{i_0}(\widetilde{\sigma}) \frac{\sum_{i,j} \pi_i \delta_{ij} \zeta_i^\alpha \lambda_i^\alpha}{\sum_i \pi_i P_{ij} \lambda_i \xi_{ij}},$$

again using Walds identity for MAPs to obtain $\mathbb{E}_{i_0}(\widetilde{\sigma}) = \mathbb{E}_{i_0}(\sigma) \sum_{i\in E} \lambda_i \pi_i P_{ij} \xi_{ij}$. $\diamond$

In the context of this thesis it is worth discussing a special case of the previous example.

Example 3.3.7 Assume that $(J_n)_{n\geq 0}$ is a discrete time semi-Markov chain on E with duration distributions $(G_{ij})_{i,j}$ and that $f : E \to \mathbb{R}$ is some function. Then consider the additive process $S_0 = 0$ and

$$S_n = \sum_{k=1}^{n} f(J_k), \quad n \geq 1$$

and its reflection, $(T_n)_{n\geq 0}$. This corresponds, in the notation of the previous example, to let $R_i = \delta_1$ and $K_i = \delta_{f(i)}$ – the Dirac measure at one and $f(i)$ respectively – in which case $L_{ij} = G_{ij}$. We get from Example 3.3.6 that if the duration distributions G_{ij} for which $f(i) > 0$ fulfill (3.24) then

$$\lim_{x\to\infty} \frac{\mathbb{P}_{i_0}(\mathcal{M}_{\widetilde{\sigma}} > x)}{\overline{G}(x)} = \mathbb{E}_{i_0}(\widetilde{\sigma}) \frac{\sum_{i,j} \pi_i \delta_{ij} f(i)^\alpha}{\sum_i \pi_i P_{ij} \xi_{ij}}.$$

Thinking of e.g. $E = \{\text{A}, \text{C}, \text{G}, \text{T}\} \times \{\text{A}, \text{C}, \text{G}, \text{T}\}$ and f some score function, such a semi-Markov model is a (probably highly unrealistic) model of DNA-sequences, for which the local similarity score can accumulate and result in a different kind of extreme value behaviour than in the *iid* case. The important message is that even with a finite state space E and a score function f taking only a finite number of different values, we can still produce a process, for which the excursions show a heavy tailed behaviour. ◇

3.4 Discussion

The assumption that $H \in \mathscr{S}^*$ for Theorem 3.2.1 to hold may seem as a rather technical requirement, which just serves to make the proofs work. However, results achieved by Foss & Zachary (2003) for the random walk with heavy tails suggest that the assumption is also necessary.

Clearly in Example 3.3.3 and Example 3.3.7 we could apply Theorem 2.4.3 if the H_{ij}-distributions are all light tailed to obtain light tailed results for fluid models controlled by a Markov chain. We have chosen to focus on the heavy tailed case. We have also chosen to state just the asymptotic tail behaviour over a regenerative cycle, since the derivation of the asymptotic extreme value distribution for the running maximum of the reflected process is then an easy consequence.

Example 3.3.7 clearly lacks some generality in the sense that we need the K_i's concentrated either on $(-\infty, 0]$ or on $(0, \infty)$. It is expected that the same result as derived in the example holds if just $\zeta_i \leq 0$ on E^- and $\zeta_i > 0$ on E^+ – still under the same moment condition imposed on K_i and R_i. To prove this we need to generalise Lemma 3.3.4, in which case we obtain from the general theory the lower bound on $\mathbb{P}_{i_0}(\mathcal{M}_{\widetilde{\sigma}} > x)$, and we need to establish a corresponding upper bound.

Finally, we could ask for more general reflected processes controlled by a semi-Markov chain. For instance, we could allow the 'rates' to be general stochastic measures on the intervals $[0, \rho_{n+1} - \rho_n)$, instead of what we essentially assume, that they are stochastic measures that are absolutely continuous w.r.t. the Lebesgue measure. This would e.g. allow for processes with jumps to be analysed. Most notably such a

generalisation includes processes modelling the residual waiting time in queueing systems with costumer arrivals being controlled by an underlying semi-Markov process. The queueing models fall outside the scope of Lemma 3.3.2, and more sophisticated tools must be developed to handle these models completely.

Notes

The results of this chapter are new – except Lemma 3.2.5, which can be found in e.g. Alsmeyer (1994) in an even more general version. The results were developed by this author together with Anders Tolver Jensen, and most of them can also be found in Hansen & Jensen (2003). The main inspiration for the proof of Theorem 3.2.1 is Asmussen (1998), who deals with heavy tailed random walks. The examples discussed in Section 3.3 are inspired by for instance Heath et al. (1997), who deal with corresponding models without the controlling semi-Markov chain. Regarding the two auxiliary lemmas, Lemma 3.3.4 and Lemma 3.3.5, this author is not convinced that they are actually new results. Even if they are not new, the proofs – especially of Lemma 3.3.4 – are probabilistic of nature and not based on analytic transform results, which may be of independent interest.

Part II

Local Alignment and Structure

4

A Brief Survey of Sequence Alignment and Structure

4.1 Introduction

We give in this chapter a survey of some biological sequence models. The models are simple and the approach is not an attempt to build completely realistic models of biological sequences. What we attempt is to give a framework for discussing alignments and structures with an emphasis on models rather than methods and algorithms. Within this framework we can derive natural test statistics for two sequences to be related or for one sequence to contain a certain structure. The main purpose is to clarify what kind of hypotheses we think of as producing 'random' sequences and what kind of alternatives we consider as producing 'non-random' sequences.

4.2 Similarity of sequences

Consider two sequences[1] $\mathbf{x} = x_1 \dots x_n$ and $\mathbf{y} = y_1 \dots y_m$ from a finite set E. We call E the alphabet and elements in E are called letters. We ask if the sequences $\mathbf{x}$ and $\mathbf{y}$ are somehow similar? How do we determine that? We will in this chapter discuss a statistical method to measure similarity and the major topic of the following chapters

[1]The word 'sequence' is the common word in probability theory, but one could argue that the computer science word 'string' is more appropriate. If not for other reasons then because we will be particularly interested in substrings, which are contiguous parts of a string, whereas a subsequence is not a contiguous part of a sequence. We will stick to sequences and use the phrase 'part of a sequence' to denote a contiguous subsequence.

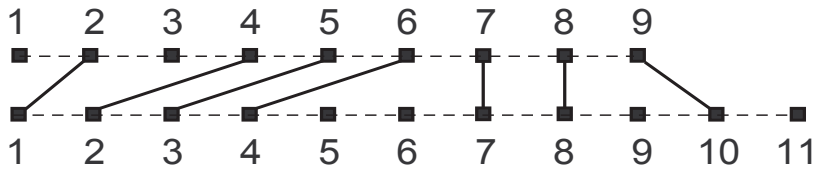

Figure 4.1: The bipartite graph representing an alignment of two sequences of length $n = 9$ and $m = 11$.

is to determine whether some measured similarity is significant. To define similarity we need to introduce the concept of an alignment.

Definition 4.2.1 *For n and m given integers we define an alignment of sequences of length n and m as a set*

$$A \subseteq \{(i,j) \mid 1 \leq i \leq n, 1 \leq j \leq m\}$$

of pairs such that for $(i,j),(i',j') \in A$ then $i < i'$ if and only if $j < j'$. For given sequences $\mathbf{x}$ and $\mathbf{y}$ of length n and m and an alignment A, the pairs $(x_i, y_j)_{(i,j)\in A}$ are called the aligned pairs and the rest of the letters are called unaligned. Furthermore, for A an alignment,

$$A_1 = \{i \mid \exists j : (i,j) \in A\}$$

denotes the 'projection' onto the first coordinate and similarly A_2 denotes the 'projection' onto the second coordinate. Finally, let $\mathcal{A}$ denote the set of all alignments of sequences of length n and m.

It may be useful to write an alignment as $A = \{(i_1, j_1), \ldots, (j_l, i_l)\}$ with $l \leq \min(n,m)$, $i_1 < \ldots < i_l$ and $j_1 < \ldots < j_l$. Figure 4.1 shows a graphical representation of an alignment as a bipartite graph.

To define the quality of an alignment we take a statistical point of view. Assume that a *null hypothesis* is given by a probability measure λ_0 on the set $E^n \times E^m$, and assume, for each alignment $A \in \mathcal{A}$, that an *alternative hypothesis* is given by a probability measure ν_A on $E^n \times E^m$. Then we will score a given alignment, A, of $\mathbf{x}$ and $\mathbf{y}$ by the log likelihood ratio $\log \nu_A(\mathbf{x},\mathbf{y}) - \log \lambda_0(\mathbf{x},\mathbf{y})$ subject to a penalty $G(A)$ with $G : \mathcal{A} \to [0,\infty]$ some function. Thus the score of an alignment $A \in \mathcal{A}$ is

$$s(A) = \log \frac{\nu_A(\mathbf{x},\mathbf{y})}{\lambda_0(\mathbf{x},\mathbf{y})} - G(A). \tag{4.1}$$

We call this score the penalised log likelihood score – given the alignment A. Typically one searches for the alignment producing the maximal score. Several optimal alignments with maximal score may exist. We will talk about *the* optimal alignment as if it was unique anyway, indicating that it is not the actual alignment with maximal score, but the maximal score itself that will attract our interest. Fast algorithms for finding the maximal score and optimal alignment(s) are of course essential for practical applications, but they will play a minor role in this thesis. We refer to Waterman (1995) for further details on algorithms.

4.2.1 Models for sequences

We will study two classes of models that provide convenient candidates for λ_0 and ν_A. We refer to these models as the independence or *iid* model and the Markov model respectively.

The independence or *iid* model is the most common choice of model. The null hypothesis is given by two probability measures λ_1 and λ_2 on E, and we define

$$\lambda_0(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} \lambda_1(x_i) \prod_{j=1}^{m} \lambda_2(y_j).$$

The alternative hypotheses are all given by a probability measure ν on $E \times E$, and we define, given an alignment A, the alternative as

$$\nu_A(\mathbf{x}, \mathbf{y}) = \prod_{(i,j)\in A} \nu(x_i, y_j) \prod_{i \notin A_1} \lambda_1(x_i) \prod_{j \notin A_2} \lambda_2(y_j).$$

Given A the log likelihood ratio is easily seen to be

$$Q_A(\mathbf{x}, \mathbf{y}) = \log \frac{\prod_{(i,j)\in A} \nu(x_i, y_j)}{\prod_{(i,j)\in A} \lambda_1(x_i)\lambda_2(y_j)} = \sum_{(i,j)\in A} \log \frac{\nu(x_i, y_j)}{\lambda_1(x_i)\lambda_2(y_j)}. \tag{4.2}$$

The Markov model provides an alternative to the independence model, which plays an important role in this thesis. To specify the null hypothesis let two $E \times E$ matrices of transition probabilities, P and Q, be given. Let x_0 and y_0 be two auxiliary letters from E and let λ_0 conditionally on (x_0, y_0) be

$$\lambda_0(\mathbf{x}, \mathbf{y} | x_0, y_0) = \prod_{i=1}^{n} P(x_{i-1}, x_i) \prod_{j=1}^{m} Q(y_{j-1}, y_j).$$

Thus the null hypothesis is just the hypothesis that the two sequences are independent Markov chains. Conditioning on (x_0, y_0) is done for notational convenience. To

specify the alternatives, let R be an $E^2 \times E^2$ matrix of transition probabilities. The alternative hypothesis ν_A, given the alignment A, is then defined as

$$\nu_A(\mathbf{x}, \mathbf{y}|x_0, y_0) = \prod_{(i,j)\in A} R\left((x_{i-1}, y_{j-1}), (x_i, y_j)\right) \prod_{i\notin A_1} P(x_{i-1}, x_i) \prod_{j\notin A_2} Q(y_{j-1}, y_j).$$

The conditional log likelihood ratio given A is seen to be

$$Q_A(\mathbf{x}, \mathbf{y}|x_0, y_0) = \sum_{(i,j)\in A} \log \frac{R\left((x_{i-1}, y_{j-1}), (x_i, y_j)\right)}{P(x_{i-1}, x_i)Q(y_{j-1}, y_j)}.$$

One observes that even though we use a conditional log likelihood – conditional on (x_0, y_0) – Q_A does in fact not depend on (x_0, y_0) unless $i_1 = 1$ and/or $j_1 = 1$, with (i_1, j_1) the first aligned pair in A. We will usually always assume that P and Q as well as R are irreducible and aperiodic.

Both models fit into the scoring schemes described in Chapter 1. For the independence model we have $f : E \times E \to \mathbb{R}$ given by

$$f(x, y) = \log \frac{\nu(x, y)}{\lambda_1(x)\lambda_2(y)},$$

and for the Markov model we have $f : E^2 \times E^2 \to \mathbb{R}$ given by

$$f((x_1, y_1), (x_2, y_2)) = \log \frac{R\left((x_1, y_1), (x_2, y_2)\right)}{P(x_1, x_2)Q(y_1, y_2)}.$$

There is no particular claim here that the Markov model (or the *iid* model for that matter) represents evolutionary events in a reasonable way. Instead, we think of the model as a pair of glasses through which we see certain features of real alignments. Given an alignment, the model provides a probability distribution on the letters in the sequences in agreement with the alignment. In that way we can, for a given dataset consisting of a number of already aligned sequences, fit a Markov model to the data and use this model to distinguish alignable sequences from non-alignable. Thus we choose to focus on features that can be captured by the Markov model, which may not be adequate in all respects, but it is a compromise between what we can analyse and implement in practice and what is a desirable trustworthy model of evolution. One argument for being interested in a Markov model versus the *iid* model – besides the fact that it is just more general – is that it makes it possible to model local (nearest neighbour) compensating mutations. That is, the change of two (or more) neighbour amino acids in a protein may be more (or less) probable than the change of one amino acid at the time. Note that this kind of modelling is *not* capable of capturing long range compensating mutations that occur to preserve the structure of proteins, say.

4.2.2 The penalty function – global and local alignment

Let an alignment $A = \{(i_1, j_1), \ldots, (j_l, i_l)\} \in \mathcal{A}$ be given. The *internal gaps* of the alignment are defined as the distances $i_{k-1} - i_k - 1$ and $j_{k-1} - j_k - 1$ between consecutive positions in the alignment. We define

$$\begin{aligned} \gamma_I(A) &= \{i_k - i_{k-1} - 1 \mid i_k - i_{k-1} > 1,\ 2 \leq k \leq l\} \\ &\quad \cup \{j_k - j_{k-1} - 1 \mid j_k - j_{k-1} > 1,\ 2 \leq k \leq l\} \end{aligned}$$

as the set of internal gaps. The set of *external gaps* $\gamma_E(A)$ consists of

$$\{i_1 - 1 \mid i_1 > 1\} \cup \{n - i_l \mid n > i_l\} \cup \{j_1 - 1 \mid j_1 > 1\} \cup \{m - j_l \mid m > j_l\}.$$

With $g_E, g_I : \mathbb{N} \to [0, \infty]$ so-called *gap penalty functions*,

$$G(A) = \sum_{k \in \gamma_E(A)} g_E(k) + \sum_{k \in \gamma_I(A)} g_I(k)$$

is a common choice of alignment penalty. It depends on the alignment only through the lengths of the gaps between aligned letters. Usually either $g_E = g_I$ or $g_E \equiv 0$, and typically

$$g_I(n) = \alpha n \quad \text{or} \quad g_I(n) = \alpha n + \beta$$

for $\alpha, \beta > 0$ some parameters. We allow for gap penalty functions to attain the value ∞, which in practice corresponds to ruling out certain alignments from consideration.

We distinguish between penalising only internal gaps ($g_E \equiv 0$) as opposed to penalising internal as well as external gaps. We refer to the former situation as *local* alignment and the latter as *global* alignment. The consequence of penalising only internal gaps is often huge when we search for the optimal alignment. And if we choose the gap penalty function properly, the local alignments will have a truly local nature, i.e. only small parts of the two sequences are aligned and the unaligned letters at the ends of the sequences are ignored. If we don't penalise the gaps sufficiently, this argument breaks down. Gaps can then be inserted in random sequences in such ways that the score will increase linearly with the length of the sequences making even the local optimal alignment close to being actually globally optimal. In fact, for random sequences there is a phase-transition behaviour in the 'penalty function space' between those penalty functions producing truly local alignments and those producing actually global alignments (Arratia & Waterman 1994).

In this thesis only local alignments will be considered, i.e. we will penalise only internal gaps. In fact, most of the theory will be developed in a limiting case for

local alignments with $g_I \equiv \infty$. For local alignments with $g_E \equiv 0$, $g_I \equiv \infty$ simply means that internal gaps are prohibited, and finding the optimal local alignment of $\mathbf{x}$ and $\mathbf{y}$ boils down to finding the best matching parts of $\mathbf{x}$ and $\mathbf{y}$ using the likelihood ratio score. This is the subject of analysis in Chapter 5 under the Markov model of the previous section.

4.3 Structure of sequences

As discussed in the introduction of the thesis, biological sequences are in fact complicated molecules with a three-dimensional structure of paramount importance for the function of the molecule. Predicting protein structure from the sequence of amino acids is a holy grail and is not possible in general with present days technology. Of a simpler nature is the so-called secondary structure of RNA, and several algorithms and fast implementations are available today, which can predict this structure directly from sequence, cf. Zuker (2003). Of course some assumptions are imposed and some limitations are still present, but the problem is nevertheless much simpler. First we need to define what we mean by a secondary structure of a sequence $\mathbf{x}$ of length n from a finite alphabet E.

Definition 4.3.1 *Given an integer n, a secondary structure, or just a structure, of a sequence of length n is a set*

$$S \subseteq \{(i,j) \mid 1 \leq i < j \leq n\}$$

of pairs such that for $(i,j),(i',j') \in S$ then $i = i'$ implies $j = j'$ and $i < i'$ implies that either $j < i'$ or $j' < j$. Let $\mathcal{S}$ denote the set of all structures of sequences of length n

One should think of a structure S as representing a two-dimensional arrangement of the sequence such that pairs in S constitute the connected or at least physically opposing letters present in the arrangement. The restriction on the pairs in a structure S implies first of all that no letter can enter two structural pairs (no triple or higher order interaction), and, secondly, two pairs are either *nested*, $i < i' < j' < j$, or *separated*, $i < j < i' < j'$. Overlapping pairs, $i < i' < j < j'$, which correspond to structural features known as *pseudoknots*, are *not* allowed. These restrictions imply that the structure can be represented as planar circular graph as shown in Figure 4.3.

The usual interpretation of a structure is chemical/physical. The RNA-molecule possesses 'unbreakable' (covalent) bonds between the consecutive letters in the sequence forming the so-called backbone, whereas the pairs in S correspond to weaker

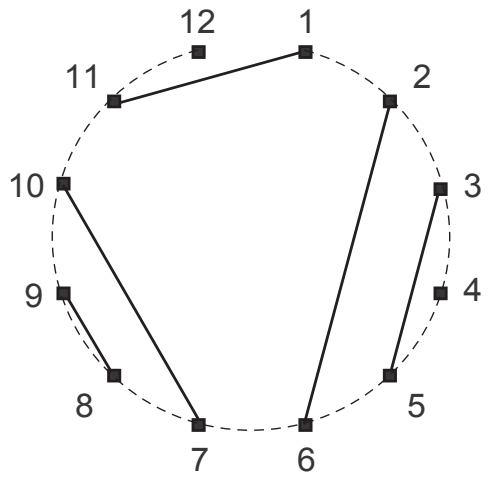

Figure 4.2: The planar circular graph representing the structure $\{(1,11),(2,6),(3,5),(7,10),(8,9)\}$ of a sequence of length $n = 12$. The dotted line represents the backbone of the structure.

hydrogen bonds. Typically, hydrogen bonds are formed in RNA between A and U and between C and G, which are the so-called Watson-Crick pairs. Hydrogen bonds are energetically favourable but since RNA is a dynamic molecule subject to the forces of the surrounding molecules (e.g. water), there may not be a unique structure. Many closely related structures may be almost equally favourable and even in some cases unrelated structures may show up to be equally favourable. We will not go into a deep discussion of structures and dynamics of large molecules in a solvent, which is far from a trivial subject. Our goal will also be quite different from actually trying to predict a structure from the sequence. Rather, we are interested in locating a part of a long sequence containing a letter composition suitable for forming specific kinds of structure. In fact, from this point of view it may not be beneficial to think of pairs in a structure as only the hydrogen bonded pairs. Instead, we may think of the pairs as the letters that are simply physically opposing each other in the structure. We continue this discussion in Chapter 7.

As for alignments we will measure the quality of a structure from a statistical point of view. Thus we will assume a *null hypothesis* given by a probability measure λ_0 on E^n, and for each structure $S \in \mathcal{S}$ an *alternative hypothesis* given by a probability measure ν_S on E^n. For a sequence $\mathbf{x}$, the score of a given structure S is then the log-likelihood ratio

$$s(S) = \log \frac{\lambda_0(\mathbf{x})}{\nu_S(\mathbf{x})} - G(S)$$

subject to the penalty given by $G : \mathcal{S} \to [0, \infty]$.

4.3.1 Special structures, models and penalties

A structure is often described using various names for substructural parts. We will not go into a detailed discussion of a classification of substructures, but restrict our attention completely to a central substructure – the so-called stems or stem-loops. In that case structures become formally very similar to alignments. As mentioned, stems are central substructural parts of all structures, and examples are considered in Chapter 7 where stem structures occur with interesting and important properties.

Definition 4.3.2 *A stem is a structure* $S = \{(i_1, j_1), \ldots, (i_l, j_l)\}$ *for which*

$$i_1 < \ldots < i_l < j_l < \ldots < j_1.$$

A stem may contain loops of various types often classified according to the following scheme:

- *The set* $i_l + 1, \ldots, j_l - 1$ *is called the hairpin loop of the stem.*
- *The coordinates* $< i_1$ *and* $> j_l$ *are sometimes called the exterior or external loop.*
- *In case* $i_k - i_{k-1} > 1$ *and* $j_{k-1} - j_k > 1$ *this is called an internal loop.*
- *In case* $i_k - i_{k-1} > 1$ *or* $j_{k-1} - j_k > 1$ *but not both this is called a bulge.*

For later use, introduce for the structure S *the sets*

$$\begin{aligned} S_1 &= \{i_1, i_2, \ldots, i_l\} \\ S_2 &= \{j_l, j_{l+1}, \ldots, j_1\} \\ H &= \{i_l + 1, \ldots, j_l - 1\} \end{aligned}$$

consisting of those coordinates that enter the structure in the first coordinate, the second coordinate, and in the hairpin-loop respectively.

Often a stem is called a stem-loop referring to the hairpin loop of the stem structure. Figure 4.3 shows a graph of a stem with one internal loop, one bulge and a hairpin loop containing four coordinates. In most real applications, a lower bound, usually 3, on the size of hairpin loops is employed, as it is difficult for the backbone to fold in a very sharp U-turn. We assume – mostly for notational reasons – that the hairpin loop size is ≥ 1, but otherwise ignore the problem.

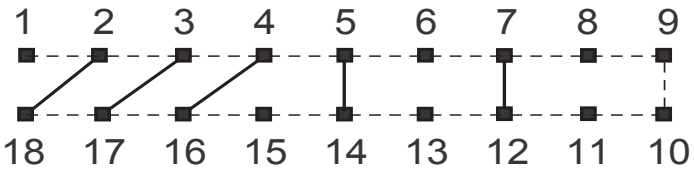

Figure 4.3: The graph representing an stem secondary structure of a sequence of length $n = 18$. The dotted line represents the backbone. The stem has a hairpin loop consisting of $8, 9, 10$, and 11, one internal loop consisting of 6 and 13 and one bulge consisting of 15. The external loop consists of 1. Note the formal relation with the graph of an alignment.

The independence or *iid* model for stems is given by a probability measure λ on E specifying the null hypothesis by

$$\lambda_0(\mathbf{x}) = \prod_{i=1}^{n} \lambda(x_i),$$

and a probability measure ν on E^2 specifying the alternative, for $S \in \mathcal{S}$ a given alignment, by

$$\nu_S(\mathbf{x}) = \prod_{i \notin S_1 \cup S_2} \lambda(x_i) \prod_{(i,j) \in S} \nu(x_j, x_i).$$

The log likelihood ratio for given S is seen to be

$$Q_S(\mathbf{x}) = \log \frac{\prod_{(i,j)\in S} \nu(x_j, x_i)}{\prod_{(i,j)\in S} \lambda(x_j)\lambda(x_i)} = \sum_{(i,j)\in S} \log \frac{\nu(x_j, x_i)}{\lambda(x_j)\lambda(x_i)}.$$

There is a deliberate interchange of the letters in (x_i, x_j) under the alternative hypothesis. This is done for comparability reasons with the Markov model introduced below, where it is most natural to consider the structure from the hairpin-loop side under the alternative hypotheses. It corresponds in some sense to a 180 degrees rotation of the stem-loop in Figure 4.3.

The Markov model for stems is defined as follows. The null hypothesis is given by an irreducible, aperiodic $E \times E$ matrix P of transition probabilities with invariant measure π_P, such that

$$\lambda_0(\mathbf{x}) = \pi(x_1) \prod_{i=2}^{n} P(x_{i-1}, x_i).$$

Thus under the null hypothesis, the sequence is a stationary Markov chain. Before introducing the alternative, it is beneficial to mentally rotate the structure 180 degrees. That is, think of the hairpin loop $x_{i_l+1}, \ldots, x_{j_l-1}$ as given. Then conditionally on $x_{i_l+1}, \ldots, x_{j_l-1}$ the sequences $x_{j_l}, x_{j_l+1}, \ldots, x_n$ and $x_{i_l}, x_{i_l-1}, \ldots, x_1$ form under the null hypothesis two independent Markov chains with transition probabilities P and $\overleftarrow{P}$ respectively. Recall the definition of the time reversed transition probabilities

$$\overleftarrow{P}_{x,y} = \frac{\pi_y P_{y,x}}{\pi_x}.$$

Thus we can rewrite $\lambda_0(\mathbf{x})$ as

$$\lambda_0(\mathbf{x}) = \pi_P(x_{i_l+1}) \prod_{i=i_l+2}^{j_l-1} P(x_{i-1}, x_i) \prod_{j=j_l}^{n} P(x_{j-1}, x_j) \prod_{i=i_l}^{1} \overleftarrow{P}(x_{i+1}, x_i)$$

The alternative is then given by an $E^2 \times E^2$ matrix R of transition probabilities, such that for a given structure S

$$\begin{aligned} \nu_S(\mathbf{x}) &= \pi_P(x_{i_l+1}) \prod_{i=i_l+2}^{j_l-1} P(x_{i-1}, x_i) \prod_{(i,j)\in S} R((x_{j-1}, x_{i+1}), (x_j, x_i)) \\ &\quad \times \prod_{i \notin S_1 \cup H} \overleftarrow{P}(x_{i+1}, x_i) \prod_{j \notin S_2 \cup H} P(x_{j-1}, x_j) \end{aligned}$$

Under the alternative, the letters in the structure arise from Markov chain transitions with bivariate transition probabilities given by R. The letters the hairpin-loop, bulges, internal loops and the external loop are conditionally independent given the other letters with the same Markov chain transition probabilities as under the null hypothesis. The log likelihood ratio given S is then

$$Q_S(\mathbf{x}) = \sum_{(i,j)\in S} \log \frac{R((x_{j-1}, x_{i+1}), (x_j, x_i))}{P(x_{j-1}, x_j)\overleftarrow{P}(x_{i+1}, x_i)}$$

As for alignments, this fits into the type of scoring schemes described in Chapter 1 with $f : E^2 \times E^2 \to \mathbb{R}$ given by

$$f((x_1, y_1), (x_2, y_2)) = \log \frac{R((x_1, y_1), (x_2, y_2))}{P(x_1, x_2)\overleftarrow{P}(y_1, y_2)}.$$

One can again discuss the rationale behind the Markov model. The reasoning is much the same as for alignments. We view the model as mostly being a suitable family of probability distributions on the sequence of letters present in a given stem-loop

structure. Note that the stem-loop Markov model can capture long range as well as nearest neighbour interactions in the sequence that occur due to the presence of structure.

In this structural setup we can define three different sets of gaps; the hairpin-loop, γ_H, the set of exterior gaps, γ_E, and the set of internal loops and bulges, γ_I:

$$\begin{aligned}
\gamma_H &= \{j_l - i_l - 1\}, \\
\gamma_E &= \{i_1 - 1 \mid i_1 > 1\} \cup \{n - j_1 \mid n > j_1\} \quad \text{and} \\
\gamma_I &= \{i_k - i_{k-1} - 1 \mid i_k - i_{k-1} > 1,\ 2 \leq k \leq l\} \\
&\quad \cup \{j_k - j_{k-1} - 1 \mid j_k - j_{k-1} > 1,\ 2 \leq k \leq l\}.
\end{aligned}$$

External, internal and hairpin-loop penalty functions g_E, g_I and g_H can then be chosen. As for alignments we distinguish between *local* stem-loops for which $g_E \equiv 0$ and *global* stem-loops otherwise. We will by theoretical methods be able to analyse local stem-loops under the Markov model in Chapter 6 in the limiting case $g_I \equiv \infty$ and $g_H \equiv 0$.

4.4 Test and classification procedures

By the formal introduction of models and hypotheses we get a very clear picture of what we test if we test the null hypothesis against the alternatives. We have discussed two different but formally very similar problems. Given two sequences do they share some similarity, i.e. does there exist a suitable alignment of the sequences. And given one sequence does it contain a structure, i.e. does there exist a suitable stem-loop structure within the sequence. Both problems are dealt with by introducing a null hypothesis and a family of alternatives – each alternative specifying a specific probability measure on the sequence(s) for a given alignment or structure. And we score each alternative using the penalised likelihood ratio. Given an optimal alignment (structure) with score s, say, the probability

$$\alpha(s) = \mathbb{P}_{\lambda_0}(\max_{A \in \mathcal{A}} s(A) > s) \quad \big(\alpha(s) = \mathbb{P}_{\lambda_0}(\max_{S \in \mathcal{S}} s(S) > s)\big)$$

is the probability under the null hypothesis of obtaining an alignment (structure) with a score $> s$, i.e. it is the test probability of testing the null hypothesis against the alternatives using the test statistic $\max_A s(A)$ (or $\max_S s(S)$).

Basically we can regard the test as a classification problem. Do the sequences show similarity or not? Does the sequence contain a stem-loop structure or not? Choosing a significance level α we can thus classify according to whether $\alpha(s) \geq \alpha$ or $\alpha(s) < \alpha$. If we compute s_0 such that $\alpha(s_0) = \alpha$ (or as close to as possible), and if the optimal

alignment scores $> s_0$ it is classified as being a true, non-random alignment, whereas if it scores $\leq s_0$ it is classified as being due to chance only. Using a classification terminology we can regard non-random optimal alignments as 'positives' and random optimal alignments as 'negatives', and we thus control the probability of predicting false positives to be α. Some might prefer to say that we control the *specificity* of the classification procedure to be $1 - \alpha$. As the alternative is composite, the control of the probability of predicting false negatives is more subtle, and we need to study the power function

$$\beta(A, s_0) = \mathbb{P}_{\nu_A}(\max_{A \in \mathcal{A}} s(A) > s_0)$$

of the test. In the classification terminology $\beta(A, s_0)$ is the *sensitivity* of the classification procedure under the specific alternative A. Given a probability measure, κ, on the set of alignments $\mathcal{A}$, we can take a Bayes approach and integrate out over $\mathcal{A}$ to get the Bayes sensitivity

$$\beta_\kappa(s_0) = \sum_{A \in \mathcal{A}} \beta(A, s_0)\kappa(A) = \sum_{\nu} \mathbb{P}_\nu(\max_{A \in \mathcal{A}} s(A) > s_0)t(\kappa)(\nu).$$

Here t is the transformation $A \mapsto t(A) = \nu_A$ from the set of alignments to the set of probability measures on the two sequences. Thus rather than thinking of κ as a probability on the alignments, one can think of $t(\kappa)$ a model of the typical distribution of the sequences under the alternative. In practice, having a dataset consisting of alignable sequences with corresponding maximal scores over all alignments being $s_1, \ldots, s_k$ we can define the *empirical sensitivity* by

$$\beta_\varepsilon(s_0) = \frac{1}{k} \sum_{i=1}^{k} 1(s_i > s_0).$$

This is the fraction of alignments we would recognise using the threshold s_0. We observe that this corresponds to letting $t(\kappa)$ be the empirical measure for the sequences contained in the dataset, and $\beta_\varepsilon(s_0)$ is thus the empirical Bayes sensitivity. We discuss in further details in Chapter 7 what the empirical sensitivity function, defined by $s_0 \mapsto \beta(s_0)$, can be used for.

However, the point of view of a simple classification problem is a (quite deliberate) over simplification of what we actually want to do. First of all, in the local alignment setup, we want not only to know that two sequences scoring $> s$ are classified as having an alignment, but we also want to identify the correct local alignment. Moreover, there may be several (truly different) local alignments that all reach a score $> s$, and we may want to consider each of these local alignments individually and not just the optimal local alignment. Thus another point of view is that for *each* alignment $A \in \mathcal{A}$ we classify A as either being a true local alignment or not according

to whether $s(A) > s$ or not for some s. From this point of view, making $\alpha(s)$ too small may not be desirable as this may result in many false negative predictions. Rather, it may be desirable just to control the number of false positive predictions to be moderate to reach a higher sensitivity. As it turns out, computing $\alpha(s)$ and giving a Poisson approximation of the number of (truly different) local alignments scoring $> s$ are closely related problems that can be dealt with in a unified way. Moreover, the asymptotic mean of the Poisson approximation for a threshold s_0 is given as $\xi(s_0) = -\log(1 - \alpha(s_0))$, so that for $\alpha(s_0)$ close to zero, $\xi(s_0) \simeq \alpha(s_0)$, whereas for large $\alpha(s_0)$ there is actually very little information in the specificity and $\xi(s_0)$ is much more interesting.

The computation of $\alpha(s)$ and the related Poisson approximation are the subjects of the following two chapters of this thesis. It is a difficult problem and in general unsolved. The reason that computing $\alpha(s)$ is difficult is that we need to understand the distribution of the maximum of a huge number of *dependent* stochastic variables. And even under the simplest independence model, the optimisation messes up the simplicity, making the problem difficult. All we can hope for are asymptotic results about $\alpha(s)$ for $n, m, s \to \infty$. In this thesis we deal with the direct asymptotic behaviour of $\alpha(s_{n,m})$ choosing $s_{n,m}$ such that $\alpha(s_{n,m})$ stays away from 0. In the realm of large deviation theory, where $\alpha(s_{n,m}) \to 0$, a logarithmic asymptotic behaviour of $\alpha(s_{n,m})$ may be derived in greater generality. That is, one can obtain expressions like

$$\lim_{n,m\to\infty} \frac{1}{nm} \log \alpha(s_{n,m}) = -\gamma$$

for some constant $\gamma > 0$. Considering the direct asymptotic behaviour of $\alpha(s)$ under the independence model and using an affine internal gap penalty function with parameters *depending on n and m*, results have recently been derived with and without a large deviation assumption (Siegmund & Yakir 2000, 2003). Direct asymptotic behaviour with a fixed affine internal gap penalty function is still out of reach with analytic methods, but a logarithmic asymptotic result was given by Yakir & Grossmann (2001). All results so far seem to have been proved under the independence model only. We will present results using the Markov models discussed, and since this will lead to enough difficulties even for a degenerate gap penalty function $g_I \equiv \infty$ (together with $g_H \equiv \infty$ in the structure setup), we will not try to give extensions allowing for gaps.

Notes

The material presented in this chapter can to some extend be found in the literature, see e.g. (Waterman 1995) and the references therein. However, the emphasis

on the models and the derivation of the score function from the models are rarely as explicit as presented here. Instead, a score function is introduced so that the computation of the maximal score becomes convenient. Secondly, the issue of significance is treated by introducing the null hypothesis of independent sequences of iid letters and arguing that the log likelihood ratio is an appropriate score function. It seems more appropriate from this authors point of view to set up the model first, derive an appropriate scoring scheme, treat the statistical issues and finally discuss algorithms and practical implementations. A reference more closely related to this point of view is Ewens & Grant (2001). Moreover, the modelling of stem-loop structures presented here in an analogous way as the modelling of alignments seems to be new. Secondary structure modelling has been given a lot of attention over the years (Zuker 2003), but the focus seems to be exclusively on the prediction of structure given the sequence. The point of view taken here is the complete opposite. We want to model the sequence given the structure, which can then be used to detect whether certain structural features are present in a sequence.

5

Local Alignment of Markov Chains

5.1 Introduction

In Chapter 4 we introduced the concept of a local alignment between two sequences using different gap penalty functions, and in this chapter we will study the distribution of the maximal, gapless ($g_I \equiv \infty$), local alignment score of two independent Markov chains. The approach is based on the theory for MAPs and the Poisson approximation presented in Chapter 2. We obtain a Poisson approximation of the number of essentially different local alignments with a score exceeding a level t, and from this we derive a Gumbel approximation of the maximal local alignment score. These results extend earlier results obtained by Dembo et al. (1994*b*) for independent sequences of *iid* variables.

5.2 Local gapless alignment

Let $(X_k)_{k\geq 1}$ and $(Y_k)_{k\geq 1}$ be two sequences of random variables taking values in the set E. We compare parts of one sequence with parts of the other using a score function $f : E \times E \to \mathbb{Z}$, and we define the random variables

$$S_{i,j}^{\Delta} = \sum_{k=1}^{\Delta} f(X_{i+k}, Y_{j+k}),$$

for $i, j, \Delta \geq 0$. The variable $S_{i,j}^{\Delta}$ is the local score of comparing the sequence $X_{i+1} \dots X_{i+\Delta}$ with the sequence $Y_{j+1} \dots Y_{j+\Delta}$.

Remark 5.2.1 *The score function f will be regarded as an E^2 column vector. Probability measures on E^2 will be regarded as E^2 row vectors, and for ν such an E^2 probability measure we use the functional analytic notation*

$$\nu(f) = \sum_{x,y} f(x,y)\nu(x,y)$$

to denote the mean of f under ν.

For $n \geq 1$ put

$$\mathcal{H}_n = \{(i,j,\Delta) \mid 0 \leq i \leq i+\Delta \leq n,\, 0 \leq j \leq j+\Delta \leq n\}$$

and call $(i,j,\Delta) \in \mathcal{H}_n$ an alignment. Note the one-to-one correspondence between elements in $\mathcal{H}_n$ and alignments in the notation of Chapter 4 of the form

$$\{(i+1,j+1),(i+2,j+2),\ldots,(i+\Delta,j+\Delta)\},$$

which justifies calling elements in $\mathcal{H}_n$ alignments.

We want to understand the distribution of the collection

$$\left(S_{i,j}^{\Delta}\right)_{(i,j,\Delta)\in\mathcal{H}_n}$$

of local scores over all alignments. We will in particular be interested in the distribution of $\mathcal{M}_n = \max_{(i,j,\Delta)\in\mathcal{H}_n} S_{i,j}^{\Delta}$ – the maximal local score over the set of alignments. We will also study the number, $C(t)$, say, of *essentially different* variables $S_{i,j}^{\Delta}$ exceeding some threshold $t \geq 0$. We will define essentially different precisely below, but just like counting excursions for a MAP we need to declump the excesses over t in some way.

The family of local scores are efficiently summarised in the matrix $(T_{i,j})_{0\leq i,j\leq n}$ defined as follows. For $i = 0$ or $j = 0$ let $T_{i,j} = 0$ and otherwise recursively define

$$T_{i,j} = \left(T_{i-1,j-1} + f(X_i, Y_j)\right)^+. \tag{5.1}$$

Thus the T-matrix correspond to the reflection of an additive process along each diagonal. Often the matrix (T_{ij}) is called the score matrix, and to see why it captures the relevant information about local scores, we observe using (2.11) that

$$\mathcal{M}_n = \max_{i,j}\ T_{i,j}. \tag{5.2}$$

This fact is closely related to the idea in the celebrated Smith-Waterman algorithm for computing the maximal local alignment score (Waterman 1995).

Definition 5.2.2 *An alignment $(i, j, \Delta) \in \mathcal{H}_n$ is called an excursion if*

$$T_{i,j} = 0, \quad S^{\delta}_{i,j} > 0 \text{ for } 0 < \delta < \Delta$$
$$\text{and either } S^{\Delta}_{i,j} = 0, \ i + \Delta = n \text{ or } j + \Delta = n$$

Let $\mathcal{E}_n$ denote the set of all excursions.

Note that $\mathcal{E}_n$ is a stochastic subset of $\mathcal{H}_n$. It follows from the definition of the score matrix $(T_{i,j})$ and the definition of an excursion that if $(i, j, \Delta) \in \mathcal{E}_n$ and $0 < \delta < \Delta$ then

$$T_{i+\delta,j+\delta} = S^{\delta}_{i,j}.$$

An excursion thus corresponds to a diagonal strip in the score matrix, for which the score starts at zero and then stays strictly positive along the diagonal until it either becomes zero again or it reaches the boundary of the score matrix.

The maximum over an excursion $e = (i, j, \Delta) \in \mathcal{E}_n$ is defined as

$$\mathcal{M}_e = \max_{0<\delta\leq\Delta} S^{\delta}_{i,j} = \max_{0<\delta\leq\Delta} T_{i+\delta,j+\delta}. \tag{5.3}$$

Definition 5.2.3 *The number of essentially different excesses over t is defined as*

$$C(t) = \sum_{e\in\mathcal{E}_n} 1(\mathcal{M}_e > t). \tag{5.4}$$

From (5.2) it follows that $(C(t) = 0) = (\mathcal{M}_n \leq t)$.

5.3 Alignment of independent Markov chains

Assume that the stochastic processes $(X_k)_{k\geq 1}$ and $(Y_k)_{k\geq 1}$ are independent Markov chains with transition probabilities P and Q respectively. Assume that P and Q are irreducible and aperiodic matrices with invariant left probability vectors π_P and π_Q respectively. Let $\pi = \pi_P \otimes \pi_Q$. Assume that the following non-degeneracy condition of f w.r.t. $P \otimes Q$ is fulfilled; for any $T \geq 1$ there exists a cycle $(x_1, \ldots, x_n)$ (w.r.t. P) and a cycle $(y_1, \ldots, y_n)$ (w.r.t. Q) such that

$$\sum_{k=1}^{n} f(x_i, y_i) \neq \sum_{k=1}^{n} f(x_i, y_{i+T \pmod n}). \tag{5.5}$$

In addition, the non-degeneracy condition from Section 2.1 must also be satisfied, i.e. there must exist cycles $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ such that

$$\sum_{k=1}^{n} f(x_k, y_k) > 0. \tag{5.6}$$

These non-degeneracy conditions don't look particularly nice in general, but are usually satisfied by quite trivial arguments in practice. In particular (5.5) looks nasty as it must hold for all $T \geq 1$. On the other hand, if we can just find x_1, x_2 and y_1, y_2 satisfying

$$f(x_1, y_1) + f(x_2, y_2) \neq f(x_1, y_2) + f(x_2, y_1), \tag{5.7}$$

then trivially (5.5) is fulfilled for all *odd* $T \geq 1$ if $P(x_1, x_2), P(x_2, x_1) > 0$ and $Q(y_1, y_2), Q(y_2, y_1) > 0$. This is, if x_1, x_2 and y_1, y_2 are two-cycles. If, in addition, $P(x_1, x_1), Q(y_1, y_1) > 0$ we can construct cycles of the form $x_1, x_2, \ldots, x_1, x_2, x_1$ and $y_1, y_2, \ldots, y_1, y_2, y_1$ such that (5.5) holds for all *even* $T \geq 1$ if just (5.7) holds. In particular, if P and Q contains only strictly positive entries, (5.7) is sufficient for (5.6) to hold. Compare with (A.2) and (A.3) in Appendix A.

For convenience we will also assume that both Markov chains are stationary, though this doesn't affect the results obtained. In this chapter, we denote by $\mathbb{P}$ the probability measure $\mathbb{P}_\pi$ under which $(X_n, Y_n)_{n\geq 1}$ is a stationary Markov chain with transition probabilities $P \otimes Q$. For notational convenience, we will also assume the existence of an auxiliary pair (X_0, Y_0) of stochastic variables, which is an initial state of the bivariate Markov chain.

In the framework of Chapter 2, the process $(S_n)_{n\geq 1}$ defined by $S_n = \sum_{k=1}^{n} f(X_k, Y_k)$ is a MAP with

$$H_{(x_0,y_0),(x_1,y_1)} = \delta_{f(x_1,y_1)},$$

and the underlying Markov chain having state space E^2 and transition probabilities $P \otimes Q$. Under the assumption that the mean drift is negative,

$$\mu = \pi(f) = \sum_{x,y\in E} f(x, y)\pi_P(x)\pi_Q(y) < 0, \tag{5.8}$$

there exists according to Lemma 2.3.1 a unique solution $\theta^* > 0$ to $\varphi(\theta) = 1$ where $\varphi(\theta)$ is the spectral radius of $\Phi(\theta)$;

$$\Phi(\theta)_{(x_0,y_0),(x_1,y_1)} := \exp(\theta f(x_1, y_1))P_{x_0,x_1}Q_{y_0,y_1}.$$

Likewise, we let K^* be the constant defined for this MAP by (2.19) in Theorem 2.4.3.

Furthermore, consider the two MAPs with the Markov chain having state space E^3 and transition probabilities $P \otimes Q \otimes Q$ and $P \otimes P \otimes Q$ respectively, and where

$$H_{(x_0,y_0,z_0),(x_1,y_1,z_1)} = \delta_{f(x_1,y_1)+f(x_1,z_1)} \quad \text{and} \quad H_{(x_0,w_0,y_0),(x_1,w_1,y_1)} = \delta_{f(x_1,y_1)+f(w_1,y_1)}.$$

These two MAPs give rise to the matrices $\Phi_i(\theta)$ for $i = 1, 2$ defined by

$$\begin{aligned} \Phi_1(\theta)_{(x_0,y_0,z_0),(x_1,y_1,z_1)} &= \exp(\theta f(x_1, y_1) + \theta f(x_1, z_1)) P_{x_0,x_1} Q_{y_0,y_1} Q_{z_0,z_1} \\ \Phi_2(\theta)_{(x_0,w_0,y_0),(x_1,w_1,y_1)} &= \exp(\theta f(x_1, y_1) + \theta f(w_1, y_1)) P_{x_0,x_1} P_{w_0,w_1} Q_{y_0,y_1}, \end{aligned}$$

and the corresponding Perron-Frobenious eigenvalues $\varphi_i(\theta)$, $i = 1, 2$.

Theorem 5.3.1 *Assume that $\mu < 0$ and that θ^* and K^* are chosen as described above. Assume, furthermore, that*

$$\varphi_1\left(\frac{3}{4}\theta^*\right) < 1 \quad \text{and} \quad \varphi_2\left(\frac{3}{4}\theta^*\right) < 1. \tag{5.9}$$

Then if we for $x \in \mathbb{R}$ define

$$t_n = \frac{\log K^* + \log n^2 + x}{\theta^*} \tag{5.10}$$

and $x_n \in [0, \theta^*)$ by $x_n = \theta^*(t_n - \lfloor t_n \rfloor)$, it holds that

$$||\mathcal{D}(C(t_n)) - \mathrm{Poi}(\exp(-x + x_n))|| \to 0 \tag{5.11}$$

for $n \to \infty$. In particular

$$\mathbb{P}(\mathcal{M}_n \leq t_n) - \exp(-\exp(-x + x_n)) \to 0 \tag{5.12}$$

for $n \to \infty$.

Remark 5.3.2 *The choice of $x_n = \theta^*(t_n - \lfloor t_n \rfloor)$ assures that $t_n - x_n/\theta^* = \lfloor t_n \rfloor \in \mathbb{Z}$. Due to the lattice effect arising from f taking values in $\mathbb{Z}$, it follows that*

$$(C(t_n) = m) = (C(t_n - x_n/\theta^*) = m)$$

as well as

$$(\mathcal{M}_n \leq t_n) = (\mathcal{M}_n \leq t_n - x_n/\theta^*),$$

and this is the reason that we need to correct by x_n is the asymptotic formulas.

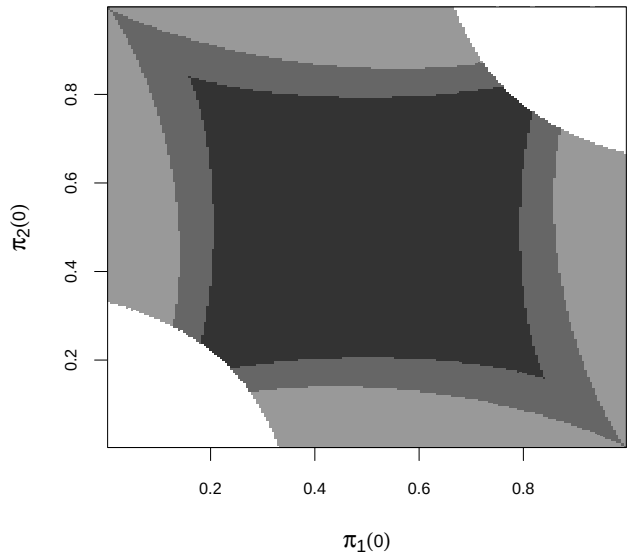

Figure 5.1: Using the score function $f(0,0) = f(1,1) = 1$ and $f(1,0) = f(0,1) = -2$, we see, as a function of $(\pi_1(0), \pi_2(0))$, the following picture: White area; $\mu \geq 0$. Light gray; only $\mu < 0$. Darker gray; in addition to $\mu < 0$, condition (5.13) is fulfilled. Darkest gray; all three conditions, $\mu < 0$, (5.13) and (5.9) are fulfilled.

The proof of Theorem 5.3.1 is not straight forward. First of all, the counting variable $C(t_n)$ is not suitable for a direct proof of the Poisson approximation. Hence we have to take a detour around another counting variable to be defined in Section 5.4. Second, to control the dependencies arising we need to establish several large deviation results, which are not all trivial, and for which we need Condition (5.9). One can show that for $i = 1, 2$ we have $\varphi_i(\theta) \leq 1$ for $0 \leq \theta \leq \theta^*/2$ but this does not extend to $\theta > \theta^*/2$ in general. Thus (5.9) is really a condition. To put some intuition into this condition, observe that φ_1 and φ_2 arise from MAPs corresponding to comparing two independent Markov chains with a third, one by one, using the same score function f. Basically, condition (5.9) ensures that it is sufficiently unlikely to obtain a high score for both comparisons at the same time. This would not be the case if f was e.g. grossly asymmetric depending almost entirely on one of the sequences, cf. also Appendix A. Annoyingly, it has not been possible to find any simple and sufficient set of conditions implying (5.9), but since $\varphi_i(3/4\theta^*)$ can be computed routinely in practice when computing θ^* anyway, the condition can easily be verified.

Example 5.3.3 The Poisson approximation given in Theorem 5.3.1 in the framework of independent sequences of *iid* variables was proved by Dembo et al. (1994*b*). When comparing their result with the present theorem – and especially the different conditions imposed – some simplification of the setup in Theorem 5.3.1 is useful.

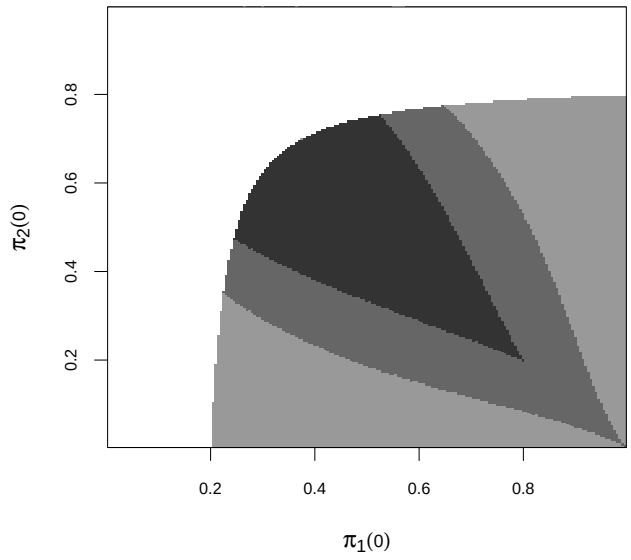

Figure 5.2: Similar to Figure 5.1 but using the asymmetric score function $f(0,0) = f(1,1) = 1,$, $f(1,0) = 0$, and $f(0,1) = -4$. White area; $\mu \geq 0$. Light gray; only $\mu < 0$. Darker gray; $\mu < 0$ and (5.13) are fulfilled. Darkest gray; $\mu < 0$, (5.13) and (5.9) are fulfilled.

Assume in this example that $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ are independent sequences of *iid* variables with the X's having distribution π_1 and the Y's having distribution π_2. Then

$$\varphi(\theta) = \mathbb{E}(\exp(\theta f(X_1, Y_1)))$$

is the Laplace transform of $\pi = \pi_1 \otimes \pi_2$, and the exponentially changed measure is given by $\pi^*(x,y) = \exp(\theta^* f(x,y))\pi_1(x)\pi_2(y)$ with $\theta^* > 0$ and $\varphi(\theta^*) = 1$. Also

$$\begin{aligned} \varphi_1(\theta) &= \mathbb{E}(\exp(\theta f(X_1, Y_1) + \theta f(X_1, Y_2))) \quad \text{and} \\ \varphi_2(\theta) &= \mathbb{E}(\exp(\theta f(X_1, Y_1) + \theta f(X_2, Y_1))) \end{aligned}$$

are Laplace transforms. The Dembo et al. (1994*b*) condition (E') for the Poisson approximation to hold can be written as

$$\pi^*(f) > 2\max\{\pi_1^*(f_1), \pi_2^*(f_2)\} \tag{5.13}$$

where π_1^* and π_2^* are the marginals of π^*,

$$\begin{aligned} f_1(x) &= \frac{1}{\theta^*}\log\sum_y \exp(\theta^* f(x,y))\pi_2(y) \quad \text{and} \\ f_2(y) &= \frac{1}{\theta^*}\log\sum_y \exp(\theta^* f(x,y))\pi_1(x). \end{aligned}$$

This condition is in fact a condition on certain relative entropies, and it relates to some large deviation properties needed in Dembo et al. (1994*b*). Our condition also relates to large deviation properties needed in the proof of Theorem 5.3.1. It is unfortunately not clear how condition (5.13) relates directly to condition (5.9). As a simple example consider $E = \{0, 1\}$,

$$\begin{aligned} f(0,0) &= f(1,1) = 1 \quad \text{and} \\ f(1,0) &= f(0,1) = -2. \end{aligned}$$

For this score function Figure 5.1 shows the set of (π_1, π_2) for which condition (5.9) as well as (5.13) are fulfilled. The figure shows that (5.9) is stronger than (5.13). We can also read of from the figure which (π_1, π_2)'s that imply $\mu < 0$ for the given score function f. A non-symmetric example is

$$\begin{aligned} f(0,0) &= f(1,1) = 1, \\ f(1,0) &= 0, \text{ and } f(0,1) = -4. \end{aligned}$$

still with $E = \{0, 1\}$. Figure 5.2 shows the set of (π_1, π_2) fulfilling $\mu < 0$, (5.13) and (5.9) respectively. In the light of Dembo et al. (1994*a*), stating that

$$\frac{\mathcal{M}_n}{\log n^2} \xrightarrow{a.s.} \frac{1}{\theta^*}$$

if and only if (5.13) is fulfilled (with $\geq$), condition (5.13) must be very close to optimal. Hence our condition must be stronger. The two examples above confirm that, but a direct proof of this has not been found. Moreover, Dembo et al. (1994*b*) find the simple and in practice relevant condition, that if $\pi_1 = \pi_2$ and f is symmetric, (5.13) is fulfilled if and only if $f(x, y)$ does not equal $f_1(x) + f_2(y)$ for all $x, y \in E$. This is unfortunately not sufficient for (5.9). ◇

5.4 The counting construction

We want $C(t)$ to be approximately Poisson distributed, but to prove that directly does not seem feasible. The size and 'shape' of $\mathcal{E}_n$ depend on the concrete realisation of the underlying stochastic variables, and the dependencies between the indicators $1(\mathcal{M}_e > t)$ seem quite complicated. The route taken is to restrict our attention to a smaller set of excursions to be defined below[1], show that the sum over this set is approximately Poisson distributed, and that the difference between that sum and $C(t)$ is asymptotically negligible.

[1] strictly speaking we will consider something slightly more complicated than just a subset of excursions

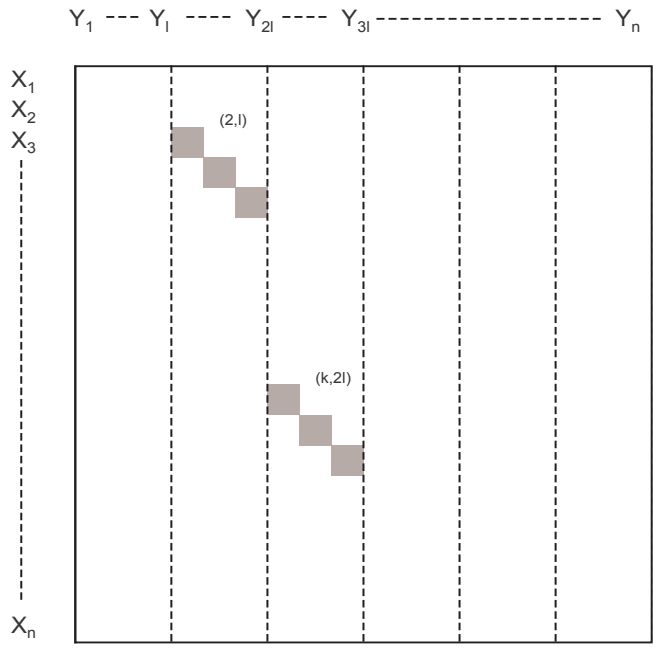

Figure 5.3: The index set is divided into vertical strips of width l – on the figure $l = 3$. Along diagonals in each strip we look for local scores exceeding the threshold. On the figure two of these diagonals-within-a-strip at position $(2, l)$ and $(k, 2l)$ respectively are shown.

We define for $l > 0$ an index set

$$I := \left\{ (k, r) = (k, ql) \mid k \in \{0, \ldots, n\},\ q \in \{0, \ldots, \left\lfloor \frac{n}{l} \right\rfloor\} \right\}.$$

One should think of this as a division of the score matrix into vertical strips of width l with I as an indexation of diagonals-within-a-strip, i.e. an index $a = (k, r) \in I$ represents the diagonal $(k + 1, r + 1), \ldots, (k + l, r + l)$, cf. Figure 5.3.

We will approximate the number of excursions exceeding t by the number of diagonals-within-a-strip containing excursions exceeding t. To be precise, we will consider, for $a = (k, r) \in I$ and $t > 0$, the variable

$$V_a \;=\; V_a(t) \;=\; 1\left(\max_{1 \le \delta \le \Delta \le l} \sum_{h=\delta}^{\Delta} f(X_{k+h}, Y_{r+h}) > t \right).$$

Thus if an excursion exceeding t is contained completely within the diagonal-within-a-strip given by a then certainly $V_a = 1$. The sum $\sum_{a \in I} V_a$ does not exactly equal the number of excursions exceeding t, though it will turn out to be a sufficiently good approximation with

$$\mathbb{P}\left(\sum_{a \in I} V_a(t_n) \neq C(t_n) \right) \to 0 \tag{5.14}$$

when $n \to \infty$ in all cases of interest.

For generality, we formulate the following result for a subset $I_0 \subseteq I$ of the index set I. This will be used in Chapter 6, cf. also the discussion at the end of this chapter. In the setup for Theorem 2.6.1, we assume for all $a \in I_0$ a subset $B_a \subseteq I_0$ given. This B_a is the neighbourhood of strong dependence of V_a, which may take various shapes in practice. Furthermore, for $a \in I_0$ let

$$\mathcal{F}_a = \sigma(V_b \mid b \notin B_a)$$

be the σ-algebra generated by those variables V_b not in the neighbourhood of strong dependence of V_a. Theorem 2.6.1 can be rephrased as:

Theorem 5.4.1 *Suppose that $l = l_n$ and $t = t_n$ chosen such that for some sequence $(\lambda_n)_{n\geq 1}$*

$$\sum_{a\in I_0} \mathbb{E}(V_a) - \lambda_n \quad \longrightarrow \quad 0, \tag{5.15}$$

for $n \to \infty$, and suppose that

$$\sum_{a\in I_0, b\in B_a} \mathbb{E}(V_a)\mathbb{E}(V_b) \quad \longrightarrow \quad 0, \tag{5.16}$$

$$\sum_{a\in I_0, b\in B_a, b\neq a} \mathbb{E}(V_a V_b) \quad \longrightarrow \quad 0, \tag{5.17}$$

$$\sum_{a\in I_0} \mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)| \quad \longrightarrow \quad 0, \tag{5.18}$$

for $n \to \infty$, then

$$\left|\left| \mathcal{D}\left(\sum_{a\in I_0} V_a\right) - \mathrm{Poi}(\lambda_n) \right|\right| \to 0. \tag{5.19}$$

In fact, the total variation norm in (5.19) is bounded by 2 times the sum of the four left hand side terms above.

In a concrete situation we need to define the B_a sets, define the sequences l and t suitably and verify condition (5.15) through (5.18). Finally we need to verify that also (5.14) is fulfilled, in which case:

Corollary 5.4.2 *If (5.19) holds and (5.14) is fulfilled too, then*

$$||\mathcal{D}\left(C(t_n)\right) - \mathrm{Poi}(\lambda_n)|| \to 0 \tag{5.20}$$

and

$$\mathbb{P}(\mathcal{M}_n \leq t_n) - \exp(-\lambda_n) \to 0. \tag{5.21}$$

5.5 Proofs

The proof of Theorem 5.3.1 is divided into a number of lemmas. We need to verify the conditions in Theorem 5.4.1, and to this end we need bounds on the expectations $\mathbb{E}(V_a V_b) = \mathbb{P}(V_a = 1, V_b = 1)$ for $a \neq b$. This is the subject of the following subsections and clearly the most difficult part of the proof. Then we collect the bounds obtained to prove that the conditions of Theorem 5.4.1 are fulfilled when aligning independent Markov chains under the assumptions given in Theorem 5.3.1. Finally we prove that (5.14) holds and the Poisson approximation of $\sum_{a\in I} V_a(t_n)$ can be translated into a Poisson approximation of $C(t_n)$.

For $a = (k, r), b = (i, j) \in I$ we always have that

$$\begin{aligned}\mathbb{E}(V_a V_b) &= \mathbb{P}(\max_{1\leq\delta\leq\Delta\leq l}\sum_{h=\delta}^{\Delta} f(X_{k+h}, Y_{r+h}) > t, \max_{1\leq\delta\leq\Delta\leq l}\sum_{h=\delta}^{\Delta} f(X_{i+h}, Y_{j+h}) > t) \\ &\leq l^4 \max_{\delta_1,\delta_2,\Delta_1,\Delta_2} \mathbb{P}(\sum_{h=\delta_1}^{\Delta_1} f(X_{k+h}, Y_{r+h}) > t, \sum_{h=\delta_2}^{\Delta_2} f(X_{i+h}, Y_{j+h}) > t). \quad (5.22)\end{aligned}$$

To bound $\mathbb{E}(V_a V_b)$ we thus need to bound the probability on the right hand side above. The same X- and/or Y- variables may enter both of the sums above in two essentially different ways. Either there are shared variables from only one of the sequences or there are shared variables from both. We will not give an exhaustive treatment of every possible ways that such a sharing of variables can take place. Rather, we treat the two essentially different cases for a specific arrangement of the sharing in sufficient details for the reader to be able to convince himself that all other arrangements can be treated similarly.

5.5.1 Positive functionals of a Markov chain

We make a useful and general observation about bounding the expectation of positive functionals, e.g. probabilities, of a Markov chain. It allows us to assume parts of the same Markov chain to be independent, stationary versions at the expense of a multiplicative constant. We call it the *decoupling argument* and state it as the following lemma:

Lemma 5.5.1 *Let $\mathcal{Z} = (Z_k)_{k\geq 0}$ be an irreducible Markov chain on a finite state space E and let $0 = k_1 < \cdots < k_N < \infty$ be given. Then there exists a constant ρ_N such that if $(Z_k^i)_{k=k_i}^{k_{i+1}}$ for $i = 1, \ldots, N$ ($k_{N+1} = \infty$) are N independent stationary Markov chains with the same transition probabilities as $\mathcal{Z}$, and $\widetilde{\mathcal{Z}} = (\widetilde{Z}_k)_{k\geq 0}$ is given*

by $\widetilde{Z}_k = Z_k^i$ *if* $k_i \leq k < k_{i+1}$ *then for a positive functional*

$$\Lambda : E^{\mathbb{N}_0} \to [0, \infty)$$

it holds that

$$\mathbb{E}(\Lambda(\mathcal{Z})) \leq \rho_N \mathbb{E}(\Lambda(\widetilde{\mathcal{Z}})). \tag{5.23}$$

The constant ρ_N *does not depend on the actual initial distribution of* $\mathcal{Z}$ *nor on the functional* Λ*.*

Proof: Assume $N = 2$. The general result follows by induction. Assume first that $\mathcal{Z}$ is stationary and that $(Z_k^1)_{k=0}^{k_2}$ and $(Z_k^2)_{k \geq k_2}$ are independent and stationary. Then $\mathcal{Z}$ has the same distribution as $\widetilde{\mathcal{Z}}$ conditionally on $Z_{k_2}^1 = Z_{k_2}^2$, hence using that Λ is a positive functional

$$\begin{aligned} \mathbb{E}(\Lambda(\mathcal{Z})) &= \frac{\mathbb{E}(\Lambda(\widetilde{\mathcal{Z}}); Z_{k_2}^1 = Z_{k_2}^2)}{\mathbb{P}(Z_{k_2}^1 = Z_{k_2}^2)} \\ &\leq \rho \mathbb{E}(\Lambda(\widetilde{\mathcal{Z}})). \end{aligned}$$

with $\rho = \left(\sum_{x \in E} \pi_x^2\right)^{-1}$, where π is the invariant distribution.

If $\mathcal{Z}$ is non-stationary with initial distribution ν, say, we have that

$$\begin{aligned} \mathbb{E}_\nu(\Lambda(\mathcal{Z})) &= \sum_{x \in E} \frac{\nu_x}{\pi_x} \pi_x \mathbb{E}_x(\Lambda(\mathcal{Z})) \\ &\leq \frac{1}{\min_x \pi_x} \mathbb{E}_\pi(\Lambda(\mathcal{Z})). \end{aligned}$$

So $\rho_2 = \rho / \min_x \pi_x$ will do. In general $\rho_N = \rho^{N-1} / \min_x \pi_x$ can be used. □

5.5.2 Variables shared in one sequence

Recall the definition of Φ, θ^* and Φ_i for $i = 1, 2$ from Section 5.3. For $\theta > 0$ denote by $r^\theta = (r^\theta(x, y))$ and $r_i^\theta = (r_i^\theta(x, y, z))$ the right Perron-Frobenious eigenvector of $\Phi(\theta)$ and $\Phi_i(\theta)$ for $i = 1, 2$ respectively. Due to irreducibility all coordinates of these vectors are strictly positive. In this section we derive a result corresponding to an overlap in the X-sequence, cf. Figure 5.4, and we thus work exclusively with the Φ_1 matrix. Similar derivations for an overlap in the Y-sequence using Φ_2 are possible.

Assume in this section that $(Z_n)_{n \geq 1}$ is a stationary Markov chain with transition probabilities Q independent of $(X_n, Y_n)_{n \geq 1}$.

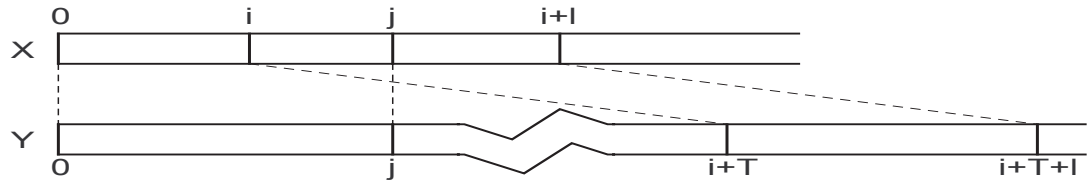

Figure 5.4: A schematic figure of how two overlapping parts of the X-sequence are compared with non-overlapping parts of the Y-sequence.

Let $i \leq j$ and $l \geq 1$ be given and define

$$\begin{aligned} S_1((x_k)_k, (y_k)_k) &= \sum_{k=1}^{i} f(x_k, y_k) \\ S_2((x_k)_k, (y_k)_k, (z_k)_k) &= \sum_{k=i+1}^{j} f(x_k, y_k) + f(x_k, z_k) \\ S_3((x_k)_k, (z_k)_k) &= \sum_{k=j+1}^{i+l} f(x_k, z_k) \end{aligned}$$

together with $S = S_1 + S_2 + S_3$.

For $\theta > 0$ define

$$\gamma^\theta = \frac{r^\theta(x_i, y_i) r_1^\theta(x_j, y_j, z_j) r^\theta(x_{i+l}, z_{i+l})}{r^\theta(x_0, y_0) r_1^\theta(x_i, y_i, z_i) r^\theta(x_j, z_j)} \exp(-i\psi(\theta) - (j-i)\psi_1(\theta) - l\psi(\theta))$$

and

$$\mathcal{L}^\theta = \gamma^\theta \exp(\theta S).$$

Note that $\gamma^\theta : E^{\mathbb{N}_0} \to [0, \infty)$ and $S : E^{\mathbb{N}_0} \to \mathbb{R}$ and hence $\mathcal{L}^\theta : E^{\mathbb{N}_0} \to [0, \infty)$ are all functionals defined on the product space $E^{\mathbb{N}_0}$.

Lemma 5.5.2 *It holds for all $\theta > 0$ that $\mathcal{L}^\theta((X_k)_k, (Y_k)_k, (Z_k)_k)$ is a likelihood, i.e.*

$$\mathbb{E}(\mathcal{L}^\theta((X_k)_k, (Y_k)_k, (Z_k)_k)) = 1, \tag{5.24}$$

and, furthermore, for $T + i \geq j$

$$\mathbb{E}(\mathcal{L}^\theta((X_k)_k, (Y_k)_k, (Y_k)_{T+k})) \leq \rho \tag{5.25}$$

for some ρ.

Proof: The first part of the lemma follows by the same arguments as those presented in Section 2.3. The only difference is that we make three different, successive exponential changes of measures. The second claim follows by the decoupling argument in Lemma 5.5.1. □

We restrict our attention to the case where $T + i \geq j$, so that there is no overlap in the Y-sequence. With abuse of notation, let S_2 denote a stochastic variable too;

$$S_2 = \sum_{k=i+1}^{j} f(X_k, Y_k) + f(X_k, Y_{T+k}).$$

Similarly, let S_1 and S_3 denote stochastic variables;

$$\begin{aligned} S_1 &= \sum_{k=1}^{i} f(X_k, Y_k), \\ S_3 &= \sum_{k=j+1}^{i+l} f(X_k, Y_{T+k}), \end{aligned}$$

and also $S = S_1 + S_2 + S_3$.

Lemma 5.5.3 *If $\theta \in (0, \theta^*]$ and $\varphi_1(\theta) \leq 1$ there exists a constant K such that*

$$\mathbb{P}(S > s) \leq K \exp(-\theta s). \tag{5.26}$$

Proof: We have that

$$\mathbb{P}(S > s) = \mathbb{E}\left(\frac{\mathcal{L}^\theta((X_k)_k, (Y_k)_k, (Y_k)_{T+k})}{\mathcal{L}^\theta((X_k)_k, (Y_k)_k, (Y_k)_{T+k})}; S > s \right).$$

The γ^θ-factor in $\mathcal{L}^\theta$ can be bounded below uniformly by b, say, since all the entries in the eigenvectors are strictly positive and since $\psi(\theta) \leq 0$ by definition of θ^* together with $\psi_1(\theta) \leq 0$ by assumption. Since we integrate over the set $(S > s)$, the exponential factor in $\mathcal{L}^\theta$ can be bounded below by $\exp(\theta s)$. Hence the denominator is bounded below by $b \exp(\theta s)$. Using (5.25) in Lemma 5.5.2 we get that

$$\mathbb{P}(S > s) \leq \rho b^{-1} \exp(-\theta s).$$

□

For variables shared in one sequence, the result in Lemma 5.5.3 is the prototypical large deviation result we can obtain using a Markov chain exponential change of measure. Of course, Lemma 5.5.3 holds for those $\theta \in (0, \theta^*]$ satisfying $\varphi_2(\theta) \leq 1$ if the overlap is in the Y sequence instead.

5.5.3 A uniform large deviation result

To handle the case with variables shared from both sequences we need a special large deviation result for Markov chains that we will derive in this section. We first state the useful Azuma-Hoeffding inequality for martingales with bounded increments, cf. Lemma 11.2 in (Waterman 1995) or Lemma 1.5 in (Ledoux & Talagrand 1991).

Lemma 5.5.4 *If $(Z_m, \mathcal{F}_m)_{m\geq 0}$ is a mean zero martingale with $Z_0 = 0$ such that for all $m \geq 1$*

$$|Z_m - Z_{m-1}| \leq c_m$$

for some sequence $(c_m)_{m\geq 1}$, then

$$\mathbb{P}(Z_m \geq \lambda) \leq \exp\left(-\frac{\lambda^2}{2\sum_{k=1}^m c_k^2}\right).$$

Fix $j \geq 1$ and let in this section $(X_k, Y_k)_{k=1}^j$ be a stationary, irreducible Markov chain with transition probabilities given by R and invariant distribution π_R. Let $(Y_k)_{k\geq j+1}$ be an independent, stationary and irreducible Markov chain with transition probabilities given by Q and invariant distribution π_Q. For an $E^2 \times E^2$-matrix G define the norm of the matrix as

$$||G||_\infty := \max_{(x,y)} \sum_{(z,w)} |G_{(x,y),(z,w)}|.$$

Since the convergence of R^k to $\mathbb{1}\pi_R$ is geometrically fast we have that

$$\sum_{k=0}^{\infty} ||R^k - \mathbb{1}\pi_R||_\infty < \infty.$$

For an E^2 vector f we let $||f||_\infty = \max_{(x,y)} |f(x,y)|$ denote the max-norm. Then clearly for any $E^2 \times E^2$ matrix G, with $G(f)$ the matrix product of G with the vector f, $||G(f)||_\infty \leq ||f||_\infty ||G||_\infty$, and especially

$$||R^k(f) - \mathbb{1}\pi_R(f)||_\infty \leq ||f||_\infty ||R^k - \mathbb{1}\pi_R||_\infty.$$

For $T \geq 1$ a fixed constant, we want to give an exponential bound of the probability

$$\mathbb{P}\left(\sum_{k=1}^j f(X_k, Y_{k+T}) \geq \sum_{k=1}^j f(X_k, Y_k)\right) \tag{5.27}$$

if $\mathbb{E}(f(X_k, Y_{k+T})) < \mathbb{E}(f(X_k, Y_k))$ all k. This is achieved by introducing a relevant martingale and then apply the Azuma-Hoeffding inequality.

Let $\mathcal{F}_m = \sigma(X_1, Y_1, \ldots, X_m, Y_m)$ ($\mathcal{F}_0 = \{\emptyset, \Omega\}$),

$$S_{j,T} = \sum_{k=1}^{j} f(X_k, Y_{k+T}) - f(X_k, Y_k) \quad (S_{0,T} = 0),$$

and with $\xi_{j,T} = \mathbb{E}(S_{j,T})$ let

$$Z_m = \mathbb{E}(S_{j,T} - \xi_{j,T} | \mathcal{F}_m). \tag{5.28}$$

Then $(Z_m, \mathcal{F}_m)_{m=0}^{j}$ is a mean zero martingale with $Z_0 = 0$ (depending on T, though we have suppressed this in the notation). Notice that $Z_j = S_{j,T} - \xi_{j,T}$. If we can bound the martingale differences

$$|Z_m - Z_{m-1}| = |\mathbb{E}(S_{j,T}|\mathcal{F}_m) - \mathbb{E}(S_{j,T}|\mathcal{F}_{m-1})|$$

with a constant, we can get the desired bound on (5.27) from the Azuma-Hoeffding inequality.

Lemma 5.5.5 *There exists a constant η independent of j and T such that*

$$|Z_m - Z_{m-1}| \leq \eta. \tag{5.29}$$

Here η can be chosen as

$$\eta = 6||f||_\infty \sum_{k=0}^{\infty} ||R^k - \mathbb{1}\pi_R||_\infty. \tag{5.30}$$

Proof: The Markov property gives that for $m \leq k \leq j$

$$\mathbb{E}(f(X_k, Y_k)|\mathcal{F}_m) = R^{k-m}(f)(X_m, Y_m),$$

and with $\overline{f}(x) = \sum_z f(x,z)\pi_Q(z)$ and $f_T(x,y) = R^T(f(x,\cdot))(x,y)$

$$\mathbb{E}(f(X_k, Y_{k+T})|\mathcal{F}_m) = \begin{cases} R^{k-m}(f_T)(X_m, Y_m) & k \in C_1 \\ R^{k-m}(\overline{f})(X_m, Y_m) & k \in C_2 \\ R^{k+T-m}(f(X_k,\cdot))(X_m, Y_m) & k \in C_3 \\ \overline{f}(X_k) & k \in C_4 \end{cases},$$

where

$$\begin{aligned} C_1 &= \{k \mid m \leq k < k+T \leq j\}, \\ C_2 &= \{k \mid m \leq k \leq j < k+T\}, \\ C_3 &= \{k \mid m-T \leq k < m \leq k+T \leq j\}, \\ C_4 &= \{k \mid m-T \leq k < m < j < k+T\}. \end{aligned}$$

Observing that

$$\mathbb{E}(S_{j,T}|\mathcal{F}_m) = \sum_{k=1}^{j} \mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_m\right) - \sum_{k=1}^{j} \mathbb{E}\left(f(X_k, Y_k)|\mathcal{F}_m\right)$$

and subtracting $\mathbb{E}(S_{j,T}|\mathcal{F}_{m-1})$ from this, the martingale difference $Z_m - Z_{m-1}$ is seen to be a sum of the following two terms

$$\begin{aligned} t_1 &= \sum_{k=m-T}^{j} \mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_m\right) - \mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_{m-1}\right) \\ t_2 &= \sum_{k=m}^{j} \mathbb{E}\left(f(X_k, Y_k)|\mathcal{F}_{m-1}\right) - \mathbb{E}\left(f(X_k, Y_k)|\mathcal{F}_m\right) \end{aligned}$$

Since

$$\begin{aligned} |\mathbb{E}\left(f(X_k, Y_k)|\mathcal{F}_m\right) - \pi_R(f)| &= |R^{k-m}(f)(X_m, Y_m) - \pi_R(f)| \\ &\leq ||f||_\infty \, ||R^{k-m} - \mathbb{1}\pi_R||_\infty, \end{aligned}$$

the term t_2 is controlled by the following inequality

$$|t_2| \leq 2||f||_\infty \sum_{k=m}^{j} ||R^{k-m} - \mathbb{1}\pi_R||_\infty \leq 2||f||_\infty \sum_{k=0}^{\infty} ||R^k - \mathbb{1}\pi_R||_\infty. \tag{5.31}$$

Noting that $||f_T||_\infty, ||\overline{f}||_\infty, ||f(x,\cdot)||_\infty \leq ||f||_\infty$ we observe that for $k \in C_1$

$$|\mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_m\right) - \pi_R(f_T)| \leq ||f||_\infty \, ||R^{k-m} - \mathbb{1}\pi_R||_\infty,$$

for $k \in C_2$

$$|\mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_m\right) - \pi_R(\overline{f})| \leq ||f||_\infty \, ||R^{k-m} - \mathbb{1}\pi_R||_\infty,$$

and for $k \in C_3$

$$|\mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_m\right) - \pi_R(f(X_k,\cdot))| \leq ||f||_\infty \, ||R^{k+T-m} - \mathbb{1}\pi_R||_\infty.$$

Since the three inequalities above also hold when conditioning on $\mathcal{F}_{m-1}$, we obtain

$$\begin{aligned} &\sum_{k \in C_1 \cup C_2 \cup C_3} |\mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_m\right) - \mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_{m-1}\right)| \\ &\quad \leq 2||f||_\infty \sum_{k \in C_1 \cup C_2} ||R^{k-m} - \mathbb{1}\pi_R||_\infty + 2||f||_\infty \sum_{k \in C_3} ||R^{k+T-m} - \mathbb{1}\pi_R||_\infty \\ &\quad \leq 4||f||_\infty \sum_{k=0}^{\infty} ||R^k - \mathbb{1}\pi_R||_\infty. \end{aligned} \tag{5.32}$$

Obviously $\mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_m\right) = \mathbb{E}\left(f(X_k, Y_{k+T})|\mathcal{F}_{m-1}\right)$ for $k \in C_4$, hence

$$|t_1| \leq 4||f||_\infty \sum_{k=0}^{\infty} ||R^k - \mathbb{1}\pi_R||_\infty,$$

which together with (5.31) gives (5.29) with η chosen as (5.30). □

Theorem 5.5.6 *If $\xi_{j,T} < 0$ it holds that*

$$\mathbb{P}(S_{j,T} \geq 0) = \mathbb{P}(S_{j,T} - \xi_{j,T} \geq -\xi_{j,T}) \leq \exp(-\frac{\xi_{j,T}^2}{2j\eta^2}) \tag{5.33}$$

with η chosen as in Lemma 5.5.5.

Proof: This follows directly from the Azuma-Hoeffding inequality for the mean zero martingale $(Z_m, \mathcal{F}_m)_{m=1}^j$, since it has increments uniformly bounded by η. □

5.5.4 Mean value inequalities

We will apply the result in the previous section by considering the Markov chain $(X_k, Y_k)_{k=1}^j$ under the exponentially tilted measure $\mathbb{P}_{\pi^*}^*$ and $(Y_k)_{k\geq j+1}$ under $\mathbb{P}_\pi$. To do so, we will need to establish inequalities relating the mean of $f(X_k, Y_k)$ to the mean of $f(X_k, Y_{k+T})$ (or $f(X_{k+T}, Y_k)$). Let in the following $\mu^* = \mathbb{E}_{\pi^*}^*(f(X_k, Y_k))$ be the stationary mean of $f(X_k, Y_k)$ under the exponentially tilted measure and let $\mu_T^* = \mathbb{E}_{\pi^*}^*(f(X_k, Y_{k+T}))$ be the stationary mean when shifting the Y-sequence T positions.

Lemma 5.5.7 *It holds that $\pi_1^* \otimes \pi_Q(f) < \mu^*$ as well as $\pi_P \otimes \pi_2^*(f) < \mu^*$.*

Proof: We consider $(X_k, Y_k)_{k\geq 1}$ under the tilted measure and an independent stationary Markov chain $(Z_k)_{k\geq 1}$ with transition probabilities Q, thus $(Z_k)_{k\geq 1}$ has the same distribution as $(Y_k)_{k\geq 1}$ does under the original measure. Then

$$(X_k, Y_k, Z_k)_{k\geq 1}$$

is a Markov chain, and we consider the MAP given by

$$\sum_{k=1}^{n} f(X_k, Z_k) - f(X_k, Y_k).$$

The Markov chain has transition probabilities

$$R^*_{(x_0,y_0,z_0),(x_1,y_1,z_1)} = \frac{r^*(x_1,y_1)}{r^*(x_0,y_0)} \exp(\theta^* f(x_1,y_1)) P_{x_0,x_1} Q_{y_0,y_1} Q_{z_0,z_1},$$

and we introduce the $\tilde{\Phi}^*(\theta)$-matrix

$$\tilde{\Phi}^*(\theta)_{(x_0,y_0,z_0),(x_1,y_1,z_1)} = \exp(\theta(f(x_1,z_1) - f(x_1,y_1))) R^*_{(x_0,y_0,z_0),(x_1,y_1,z_1)}.$$

Clearly, with $\tilde{\varphi}^*(\theta) = \mathrm{spr}(\tilde{\Phi}(\theta))$ we have that $\tilde{\varphi}^*(0) = \tilde{\varphi}^*(\theta^*) = 1$ ($\tilde{\Phi}^*(0)$ is stochastic and $\tilde{\Phi}^*(\theta^*)$ has right eigenvector $r^*(x_1,z_1)/r^*(x_1,y_1)$). Moreover, (5.5) together with (2.6) assures that $\tilde{\varphi}^*$ is strictly convex, and since

$$\partial_\theta \tilde{\varphi}^*(0) = \pi_1^* \otimes \pi_Q(f) - \mu^*$$

by (2.7), it follows that $\pi_1^* \otimes \pi_Q(f) < \mu^*$. The second inequality follows similarly. □

Lemma 5.5.8 *The sequence $(\mu_T^*)_{T\geq 1}$ is convergent and with*

$$\mu_\infty^* = \lim_{T\to\infty} \mu_T^*$$

it holds that $\mu_\infty^* < \mu^*$.

Proof: We first observe that

$$\mu_T^* = \mathbb{E}^*_{\pi^*}(f(X_1, Y_{1+T})) \to \pi_1^* \otimes \pi_2^*(f)$$

for $T \to \infty$, where π_1^* and π_2^* are the marginals of π^*.

We consider $(X_k, Y_k)_{k\geq 1}$ under the tilted measure and let $(W_k, Z_k)_{k\geq 1}$ be an *independent* copy with the same distribution. Then

$$(X_k, W_k, Y_k, Z_k)_{k\geq 1}$$

is a Markov chain and we will consider the MAP given by

$$\sum_{k=1}^{n} f(X_k, Z_k) + f(W_k, Y_k) - f(X_k, Y_k) - f(W_k, Z_k).$$

Introducing the corresponding $\Phi^*_\infty(\theta)$ matrix and its spectral radius φ^*_∞ we derive, by similar arguments as in the previous lemma, that $\varphi^*_\infty(0) = \varphi^*_\infty(\theta^*) = 1$, that $\varphi^*_\infty(\theta)$ is strictly convex, and that $\partial_\theta \varphi^*_\infty(0) = 2\mu^*_\infty - 2\mu^*$. Hence $\mu^*_\infty < \mu^*$. □

Interestingly, the inequality in previous lemma does not only hold in the limit but actually for all T.

Lemma 5.5.9 *For all $T \geq 1$ we have that*

$$\mu_T^* < \mu^*. \tag{5.34}$$

Proof: With $S_n^T = \sum_{k=1}^n f(X_k, Y_{k+T})$ and $S_n = \sum_{k=1}^n f(X_k, Y_k)$ we observe that $S_n \stackrel{\mathcal{D}}{=} S_n^T$, since under $\mathbb{P} = \mathbb{P}_\pi$ the X- and Y-sequence are independent, stationary Markov chains. We then obtain by (2.8) for $\theta > 0$ that

$$\frac{1}{n} \log \mathbb{E}(\exp(\theta S_n^T)) = \frac{1}{n} \log \mathbb{E}(\exp(\theta S_n)) \to \psi(\theta) \tag{5.35}$$

for $n \to \infty$.

Consider first the case $T = 1$ and the stacked Markov chain

$$(X_k, X_{k+1}, Y_k, Y_{k+1})_{k \geq 1},$$

which under the tilted measure has transition probabilities

$$R^*_{(x_0,w_0,y_0,z_0),(x_1,w_1,y_1,z_1)} = \frac{r^*(w_1, z_1)}{r^*(w_0, z_0)} \exp(\theta^* f(w_1, z_1)) P_{w_0,w_1} Q_{z_0,z_1} \delta_{w_0,x_1} \delta_{z_0,y_1}.$$

Introduce the matrix

$$\Phi_1^*(\theta)_{(x_0,w_0,y_0,z_0),(x_1,w_1,y_1,z_1)} = \exp(\theta(f(x_1, z_1) - f(w_1, z_1))) R^*_{(x_0,w_0,y_0,z_0),(x_1,w_1,y_1,z_1)}$$

and its spectral radius $\varphi_1^*(\theta) = \mathrm{spr}(\Phi_1^*(\theta))$. Clearly, $\varphi_1^*(0) = 1$ and we observe that

$$\Phi_1^*(\theta^*)_{(x_0,w_0,y_0,z_0),(x_1,w_1,y_1,z_1)} = \frac{r^*(w_1, z_1)}{r^*(w_0, z_0)} \exp(\theta^* f(x_1, z_1)) P_{w_0,w_1} Q_{z_0,z_1} \delta_{w_0,x_1} \delta_{z_0,y_1}.$$

The matrix $\Phi_1^*(\theta^*)$ has the same spectral radius if we remove the eigenvector fraction, and by (2.8) together with (5.35) we obtain that

$$\psi_1^*(\theta^*) = \log \varphi_1^*(\theta^*) = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}(\exp(\theta^* S_n^1)) = \psi(\theta^*) = 0,$$

thus $\varphi_1^*(\theta^*) = 1$.

Furthermore, by (2.7) $\partial_\theta \varphi_1^*(0) = \mu_1^* - \mu^*$. Using (5.5) (for $T = 1$) together with (2.6) we get that φ_1 is strictly convex, hence

$$\mu_1^* < \mu^*.$$

A similar argument for general T is possible by introducing the stacked Markov chain

$$(X_k, \ldots, X_{k+T}, Y_k, \ldots, Y_{k+T})_{k \geq 1}$$

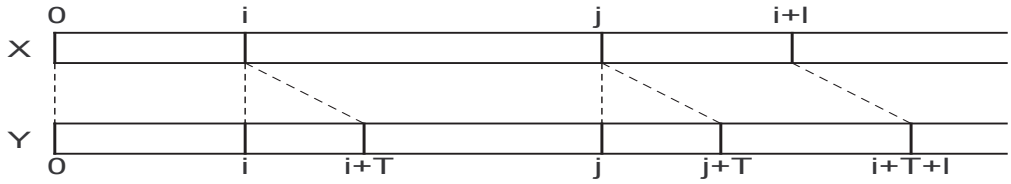

Figure 5.5: This figure shows the case with overlap in both sequence.

on $E^T \times E^T$ and the MAP given by

$$\sum_{k=1}^{n} f(X_k, Y_{k+T}) - f(X_{k+T}, Y_{k+T}).$$

The spectral radius $\varphi_T^*(\theta)$ of the corresponding matrix $\Phi_T^*(\theta)$ for this MAP fulfills that $\varphi_T^*(0) = \varphi_T^*(\theta^*) = 1$, $\partial_\theta \varphi_T^*(0) = \mu_T^* - \mu^*$ and it is strictly convex by (5.5) and (2.6). Thus $\mu_T^* < \mu^*$. □

A similar result is possible when shifting the X-sequence instead. The interpretation is rather pleasing. Under the tilted measure, the distribution of the Markov chain $(X_n, Y_n)_{n\geq 1}$ is designed so that the pairs (X_n, Y_n) 'match' well – using the score function f. If we shift one of the sequences this matching is somehow destroyed, and the lemma assures us that we can not under the tilted measure obtain an average matching by shifting that is as good as if we don't shift.

5.5.5 Variables shared in both sequences

We define for $i, j, l, T \geq 1$ with $i \leq j$

$$\begin{aligned} S_1 &= \sum_{k=1}^{i} f(X_k, Y_k), \quad S_2 = \sum_{k=i+1}^{j} f(X_k, Y_k) \\ \widetilde{S}_2 &= \sum_{k=i+1}^{j} f(X_k, Y_{k+T}), \quad S_3 = \sum_{k=j+1}^{i+l} f(X_k, Y_{k+T}), \end{aligned}$$

cf. Figure 5.5.

Theorem 5.5.10 *There exists an $\varepsilon > 0$ and some K (both independent of T) such that*

$$\mathbb{P}(S_1 + S_2 > t, \widetilde{S}_2 + S_3 > t) \leq K \exp(-\theta^*(1+\varepsilon)t) \tag{5.36}$$

for $t \geq 0$.

Proof: Assume first that the number of variables $j-i$ in the overlapping part is small, less than $t(4||f||_\infty)^{-1}$, say, in which case we obtain the estimate

$$\begin{aligned}\mathbb{P}(S_1+S_2>t,\widetilde{S}_2+S_3>t) &\leq \mathbb{P}(S_1>3/4t,S_3>3/4t)\\ &\leq \rho\mathbb{P}(S_1>3/4t)\mathbb{P}(S_3>3/4t)\\ &\leq K\exp(-3/2\theta^* t),\end{aligned}$$

using the decoupling argument for the second inequality and then a standard exponential change of measure argument. This implies (5.36) with $\varepsilon = 1/2$.

If instead $j-i \geq t(4||f||_\infty)^{-1}$ we observe that

$$\begin{aligned}&\mathbb{P}(S_1+S_2>t,\widetilde{S}_2+S_3>t)\\ &\leq \mathbb{P}(S_1+S_2>t,\widetilde{S}_2\geq S_2)+\mathbb{P}(\widetilde{S}_2+S_3>t,S_2\geq\widetilde{S}_2).\end{aligned} \tag{5.37}$$

With $L_j^{\theta^*} = r(X_j,Y_j)/r(X_0,Y_0)\exp(\theta^*(S_1+S_2))$ we obtain

$$\begin{aligned}\mathbb{P}_\pi(S_1+S_2>t,\widetilde{S}_2\geq S_2) &= \mathbb{P}_\pi\left(\frac{L_j^{\theta^*}}{L_j^{\theta^*}};S_1+S_2>t,\widetilde{S}_2\geq S_2\right)\\ &\leq b^{-1}\exp(-\theta^* t)\mathbb{P}^*_{\pi,j}(\widetilde{S}_2\geq S_2)\end{aligned}$$

where $\mathbb{P}^*_{\pi,j}$ denotes the tilted measure *up to index* j. Using the decoupling argument from Lemma 5.5.1, we can, at the expense of a factor ρ, assume that the sequence $(X_k,Y_k)_{k=i}^j$ is a stationary Markov chain under the tilted measure and that $(Y_k)_{k\geq j+1}$ is independent and stationary under the original measure. Under this assumption it follows that the mean of $\widetilde{S}_2-S_2$ equals $(j-T-i)\mu_T^*+T\pi_1^*\otimes\pi_Q(f)-(j-i)\mu^*$. Using Lemma 5.5.7, Lemma 5.5.8, and Lemma 5.5.9 we can find a $\zeta>0$, independent of T, such that

$$(j-T-i)\mu_T^*+T\pi_1^*\otimes\pi_Q(f)-(j-i)\mu^* < -(j-i)\zeta.$$

Hence Theorem 5.5.6 gives that

$$\mathbb{P}^*_{\pi,j}(\widetilde{S}_2\geq S_2)\leq\rho\exp(-\frac{\zeta^2(j-i)}{2\eta^2})\leq\rho\exp(-\frac{\zeta^2 t}{8||f||_\infty\eta^2})$$

or, with $\varepsilon=\zeta^2(\theta^*8||f||_\infty\eta^2)^{-1}$,

$$\mathbb{P}(S_1+S_2\geq t,\widetilde{S}_2\geq S_2)\leq\rho b^{-1}\exp(-\theta^*(1+\varepsilon)t)$$

and (5.36) follows. Of course, a similar argument takes care of the second term in (5.37). □

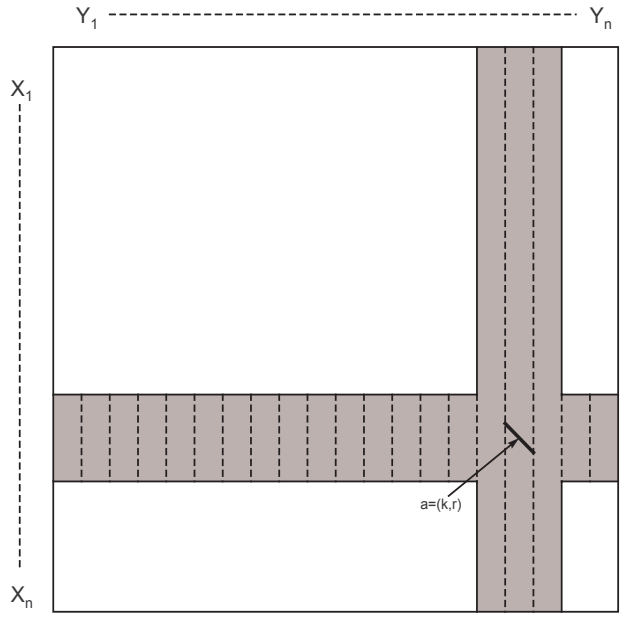

Figure 5.6: For each diagonal-within-a-strip given by a we define a neighbourhood of strong dependence. This figure shows the neighbourhood for some $a = (k, r)$. The 'arms' of this cross correspond to diagonals-within-a-strip, b, sharing variables with a from either the X- or Y-Markov chain but not both. In the intersection, b can share variables with a from both the sequences. The dashed lines mark the strips into which the matrix is divided.

5.5.6 Proof of the Poisson approximation

Returning to the alignment of Markov chains considered in this chapter, we define the neighbourhood of strong dependence B_a, for $a = (k, r) \in I$, by

$$\begin{aligned} B_a^1 &= \{k-l, \ldots, k+2l\} \times \{0, l, 2l, \ldots, \left\lfloor \frac{n}{l} \right\rfloor l\} \\ B_a^2 &= \{0, \ldots, n\} \times \{r-l, r, r+l\} \end{aligned}$$

and then $B_a = B_a^1 \cup B_a^2$. Note that $\max_a |B_a| = O(n)$. The set B_a^1 is a horizontal strip of strong dependence and B_a^2 is a vertical strip of strong dependence. Of course, the set should be properly modified close to the boundaries of the index set, and we could chose to always consider $B_a \cap I$. This boundary modification is insignificant and will be ignored throughout.

Lemma 5.5.11 *If we, for some $x \in \mathbb{R}$, let*

$$l = l_n \sim (\log n^2)^3 \quad \text{and} \quad t = t_n = \frac{\log K^* + \log n^2 + x}{\theta^*} \tag{5.38}$$

and define $x_n \in [0, \theta^*)$ by $x_n = \theta^*(t_n - \lfloor t_n \rfloor)$, then under the assumptions in Theorem 5.3.1, the conditions in Theorem 5.4.1 are fulfilled with

$$\lambda_n = \exp(-x + x_n).$$

That is

$$\left|\left|\mathcal{D}\left(\sum_{a\in I} V_a\right) - \mathrm{Poi}(\exp(-x+x_n))\right|\right| \to 0.$$

Proof: According to Lemma 2.4.6

$$\mathbb{P}(V_{(0,0)}=1) = \mathbb{P}\left(\max_{1\le\delta\le\Delta\le l}\sum_{k=\delta}^{\Delta} f(X_k,Y_k) > t - x_n/\theta^*\right) \sim l n^{-2}\exp(-x+x_n) \quad (5.39)$$

for $n\to\infty$. Since $|I|\sim n^2 l^{-1}$ and, due to stationarity, all the events $(V_a=1)$ for $a\in I$ are equally probable, it follows from (5.39) that

$$\sum_{a\in I}\mathbb{E}(V_a) = |I|\mathbb{P}\left(\max_{1\le\delta\le\Delta\le l}\sum_{k=\delta}^{\Delta} f(X_k,Y_k) > t - x_n/\theta^*\right) \sim \exp(-x+x_n),$$

or, since $\exp(-x+x_n)$ is bounded,

$$\left|\sum_{a\in I}\mathbb{E}(V_a) - \exp(-x+x_n)\right| \to 0$$

for $n\to\infty$.

Furthermore, we observe that $|I|\mathbb{E}(V_{(0,0)})$ is bounded and that $\max_a|B_a| = o(|I|)$ for $n\to\infty$, so condition (5.16) is fulfilled by

$$\sum_{a\in I, b\in B_a}\mathbb{E}(V_a)\mathbb{E}(V_b) \le |I|\times\max_a|B_a|\times\mathbb{E}(V_{(0,0)})^2 \;=\; O\left(\frac{\max_a|B_a|}{|I|}\right)\to 0.$$

We prove that (5.17) is fulfilled by splitting the set B_a into three disjoint sets and, depending on the set, give a bound of $\mathbb{E}(V_aV_b)$ for b in each of these sets. For $a\in I$ let

$$B_a = C_a\cup D_a^1\cup D_a^2$$

with C_a, D_a^1 and D_a^2 being the disjoint sets

$$C_a = B_a^1\cap B_a^2, \quad D_a^1 = B_a^1\backslash C_a \quad\text{and}\quad D_a^2 = B_a^2\backslash C_a$$

We see that C_a is the centre of the set B_a, and D_a^1 and D_a^2 are the remaining horisontal and vertical parts.

Consider the case $b\in C_a$ and $b\neq a$. Using (5.22) together with Lemma 5.5.10 we can find an $\varepsilon>0$ such that

$$\mathbb{E}(V_aV_b) \le Kl^4\exp(-\theta^*(1+\varepsilon)t).$$

Hence, observing that $\sum_{a\in I}|C_a| \le 9|I|l \le 9n^2$,

$$\sum_{a\in I, b\in C_a, b\neq a} \mathbb{E}(V_aV_b) \;\le\; Kl^4 n^{-2(1+\varepsilon)} \sum_{a\in I}|C_a| \to 0$$

for $n\to\infty$.

For $b\in D_a^i$, Lemma 5.5.3 with $\theta = 3/4\theta^*(1+\varepsilon)$ for some $\varepsilon>0$ applies due to (5.9), and together with (5.22) we obtain that

$$\mathbb{E}(V_aV_b) \le Kl^4 \exp(-3/2\theta^*(1+\varepsilon)t),$$

in which case

$$\begin{aligned}\sum_{a\in I, b\in D_a^i} \mathbb{E}(V_aV_b) &\le Kl^4 n^{-3(1+\varepsilon)} \sum_{a\in I}|D_a^i| \\ &\le Kl^4 n^{-3(1+\varepsilon)} 8n^3 \to 0.\end{aligned}$$

The two-dimensional process $(X_k, Y_k)_{k\ge 1}$ is a stationary, irreducible Markov chain on a finite state space, hence we can extend it to a doubly infinite, stationary process $(X_k, Y_k)_{k\in\mathbb{Z}}$, which is exponentially β-mixing. The β-mixing coefficients thus satisfy

$$\beta(k) \le K_1 \exp(-K_2 k)$$

for some constants $K_1, K_2 > 0$. For $a=(r,r)\in I$ we define $I_1 = (-\infty, r-l]$, $I_2 = [r+1, r+l]$, and $I_3 = [r+2l+1, \infty)$, for which $d(I_1\cup I_3, I_2) = l+1$. Then clearly with $I = I_1\cup I_3$ and $J = I_2$, $\mathcal{F}_a \subseteq \mathcal{F}_I = \sigma(X_n, Y_n \mid n\in I_1\cup I_3)$ and V_a is measurable w.r.t. $\mathcal{F}_J = \sigma(X_n, Y_n \mid n\in I_2\}$. So Lemma 2.6.2 together with Theorem 2.6.4 imply that

$$\mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)| \le 2\alpha(\mathcal{F}_I, \mathcal{F}_J) \le 2\beta(l+1) \le K\exp(-K_2 l).$$

For any non-diagonal $a=(k,r)\in I$ we can shift the X-process by stationarity to reduce the problem to the previous one and thus to obtain the same bound. This bound implies that

$$\sum_{a\in I} \mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)| \;\le\; Kn^2 \exp(-K_2(\log n^2)^3) \to 0$$

for $n\to\infty$. □

Proof of Theorem 5.3.1: We prove that

$$\mathbb{P}\left(\sum_{a\in I} V_a(t_n) \neq C(t_n)\right) \to 0. \tag{5.40}$$

We first show that the probability of (i) the existence of an excursion in $\mathcal{E}_n$ exceeding t and crossing a boundary between two diagonals-within-a-strip and (ii) two exceedances fall within the same diagonal-within-a-strip tends to zero. For $a \in I$ we introduce the excursion containing a

$$e_a = (i, j, \Delta) \quad \text{if} \quad a \in \{(i+1, j+1), \ldots, (i+\Delta, j+\Delta)\},$$

and then using Lemma 2.6.7

$$\begin{aligned} \mathbb{P}(\exists a \in I : \mathcal{M}_{e_a} > t) &\leq K(|I|(\log n)^2 \exp(-\theta^* t) + |I| n^{-2}) \\ &\leq K_1((\log n)^{-1} + (\log n)^{-3}) \to 0. \end{aligned}$$

The other way around, let $\mathcal{E}_n^a$ be the set of excursions contained in the diagonal-within-a-strip $a = (k, r) \in I$, that is, with $A = \{(k+1, r+1), \ldots, (k+l, r+l)\}$

$$e = (i, j, \Delta) \in \mathcal{E}_n^a \quad \text{if} \quad (i+1, j+1), (i+\Delta, j+\Delta) \in A.$$

Then for a given $a \in I$ we get, using (2.22) and a decoupling argument, that

$$\mathbb{P}(\exists e, e' \in \mathcal{E}_n^a : \mathcal{M}_e > t, \mathcal{M}_{e'} > t) \leq K l^3 \exp(-2\theta^* t).$$

Hence the probability that two excursions exceeding t occur within the same diagonal-within-a-strip is

$$\mathbb{P}(\exists a \in I \ \exists e, e' \in \mathcal{E}_n^a : \mathcal{M}_{e'} > t, \mathcal{M}_{e'} > t) \quad \leq \quad K l^2 n^{-2} \to 0$$

for $n \to \infty$.

Finally, there can be some problems close to the boundary of the score matrix. Introduce the set $\tilde{\mathcal{E}}_n \subseteq \mathcal{E}_{n+l}$ with $(i, j, \Delta) \in \tilde{\mathcal{E}}_n$ if

$$(i, j, \Delta) \in \mathcal{E}_{n+l} \text{ and either } i \leq l, i \geq n - l, \text{ or } j \geq n - l.$$

Thus $\tilde{\mathcal{E}}_n$ consists of those excursions that occur close to the boundary – in the enlarged $n+l \times n+l$ score matrix, cf. the definition of I, which allows for diagonals-within-a-strip to extend a little beyond the boundary of the score matrix. Clearly,

$$\mathbb{P}(\exists e \in \tilde{\mathcal{E}} : \mathcal{M}_e > t) \leq K n l \exp(-\theta^* t) \leq K_1 l n^{-1} \to 0.$$

One then just have to observe that the event $(\sum_{a \in I} V_a(t_n) \neq C(t_n))$ is contained in the union of the three events treated above, hence (5.40) holds. □

5.6 Discussion

For notational convenience it was chosen to state and prove the results in this chapter using a score function f that compares the sequences letter by letter. There is actually no loss of generality in doing so, since if we want to consider more complicated score functions depending on more letters, we can always stack the Markov chain. In particular, the use of a score function like the one derived for the Markov model in Chapter 4 is also covered by Theorem 5.3.1. In fact, Theorem 5.3.1 is substantiable more general than it looks at first. By staking the process, it follows from the theorem that when comparing *independent* sequences of n'th order Markov chains, say, using a score function $f : E^w \to E^w \to \mathbb{Z}$ with a window of size w, one can derive the same asymptotic theory for the maximal, local (gapless) alignment score as derived by Dembo et al. (1994*b*) for the *iid* case. Even in the framework of *iid* sequences the theorem provides a generalisation to score functions that compare more than one letter at the time. Of course, the conditions imposed for the theorem to hold, especially (5.9), puts some restrictions on the generality, but it is nevertheless noteworthy that the same kind of extreme value theory extend to a vary large class of processes and score functions.

It is unfortunate that we haven't been able to identify sufficient conditions for Theorem 5.3.1 that are equivalent with the condition (E') in Dembo et al. (1994*b*) in the *iid* setup. Attempts of improvements should quite clearly be targeted at Lemma 5.5.3, which is the reason for assuming (5.9). Referring to Appendix A, this result has be optimised as much as possible using the straight forward exponential change of measure. Thus to achieve improvements one much invent another approach.

One can also allow for the sequences that are compared to have different lengths m and n, say. In this case a similar result is valid for $n, m \to \infty$ with

$$t_{m,n} = \frac{\log K^* + \log(mn) + x}{\theta^*}.$$

This can be derived using $I_0 \subseteq I$ given by

$$I_0 = \left\{ (k,r) = (k,ql) \mid k \in \{0,\ldots,m\},\ q \in \{0,\ldots,\left\lfloor \frac{n}{l} \right\rfloor\} \right\},$$

and then apply Theorem 5.4.1. Some restrictions on the simultaneous growth of m and n must be made in order for this to work. It is certainly sufficient that $0 < \liminf m/n \leq \limsup m/n < \infty$ but weaker assumptions will do. From a practical point of view, one should just keep in mind that when m and n are of vastly different size there could be problems with the Poisson and the Gumbel approximations. Simulation studies can then be helpful in revealing to what extend the approximations hold anyway. We refer to Chapter 7 for a simulation study exploring a similar problem in the context of local structures as presented in Chapter 6.

Notes

*The results in this chapter are new. The general idea of using the Poisson approximation by Arratia et al. (1989) for proving a result like Theorem 5.3.1 is, on the other hand, not. A very similar technique of proof was used by Dembo et al. (1994*b*). Indeed, Arratia et al. (1989) themselves apply the Poisson approximation to a similar sequence comparison problem. But even though Dembo et al. (1994*b*) was a great inspiration for the results and proofs presented here, almost all of the details have been changed. The counting construction presented here to approximate $C(t_n)$ seems, in this authors opinion, to be more closely related to $C(t_n)$ and easier to work with than the corresponding counting construction presented by Dembo et al. (1994*b*). Moreover, they use to a great extend inequalities based on large deviation theory to derive relevant estimates of probabilities, whereas the the method of choice in this chapter is exponential change of measure. Furthermore, several new problems arise due to the Markov dependence in the sequences. Most notably, the result corresponding to Theorem 5.5.10 was derived in the iid setup by a very smart permutation argument, which essentially employs exchangeability of iid-variables. The alternative route presented here, which works for Markov chains, is an application of the Asuma-Hoeffding inequality for martingales.*

6

Local Folding of Markov Chains

6.1 Introduction

In Chapter 4 we discussed structures and the scoring of especially stem-loops. This chapter contains an analysis of the maximal, gapless ($g_I \equiv \infty$) score of local stem-loop structures, allowing in principle an arbitrarily large hairpin loop ($g_H \equiv 0$). For the sequence being a Markov chain, we obtain a Poisson approximation of the number of essentially different stem-loop structures with a score exceeding some level t, and from this we derive a Gumbel approximation of the maximal local stem-loop score.

6.2 Local folding structures

Let $(X_k)_{k\geq 1}$ be a sequence of stochastic variables taking values in a finite set E, and let, for $n \geq 1$ given, $Y_k = X_{n-k+1}$, $k = 1, \ldots, n$. Thus $(Y_k)_{1\leq k\leq n}$ is the time reversion of $(X_k)_{1\leq k\leq n}$. Introduce the set

$$\mathcal{H}_n = \{(i, j, \Delta) \mid 0 \leq j \leq j + \Delta \leq n,\, n - j + 1 \leq i \leq i + \Delta \leq n\},$$

which we in this chapter call the set of structures. As for alignments, there is a one-to-one correspondence between the elements $(i, j, \Delta) \in \mathcal{H}_n$ and structures of the form

$$\{(n - j - \Delta + 1, i + \Delta), (n - j - \Delta + 2, i + \Delta - 1), \ldots, (n - j, i + 1)\}.$$

The structures given by $\mathcal{H}_n$ are closely related to the alignments defined in Chapter 5, and the only difference is the restriction $n + 1 \leq i + j$. This reflects the physical restriction of folding a sequence back onto itself to form a stem-loop.

We assume a score function $f : E \times E \to \mathbb{Z}$ given and define for $(i, j, \Delta) \in \mathcal{H}_n$

$$S_{i,j}^{\Delta} = \sum_{k=1}^{\Delta} f(X_{k+i}, Y_{k+j}) = \sum_{k=1}^{\Delta} f(X_{k+i}, X_{n-k-j+1})$$

as the score of the structure (i, j, Δ). The score $S_{i,j}^{\Delta}$ is thus in the terminology of Chapter 4 the local stem-loop score of the structure given by (i, j, Δ) using $g_H \equiv 0$.

We can summarise the scores by the score matrix $(T_{i,j})$ – still reflecting the constraints imposed by $\mathcal{H}_n$ – defined by $T_{i,j} = 0$ for $i + j \leq n + 1$ and recursively

$$T_{i,j} = (T_{i-1,j-1} + f(X_i, Y_j))^+ \tag{6.1}$$

for $i + j > n + 1$. Note that the score matrix is only non-zero below the 'anti'-diagonal, that is, below the diagonal going from the lower left corner to the upper right corner. In is quite common in the literature on secondary structures to rotate the score matrix 90 degrees counter clockwise, so that it becomes an upper triangular matrix. We keep it this way to make it directly comparable with the alignment score matrix from the previous chapter.

We want to count the essentially different excesses above a threshold in the score matrix, but we will further impose a *window* of size w, $2 \leq w \leq n$, to which we will restrict our attention. One observes as for alignments that the maximum of $S_{i,j}^{\Delta}$ where $(i, j, \Delta) \in \mathcal{H}_n$ is restricted by the window to satisfy $i + j + 2\Delta \leq n + w$ can be obtained as the maximum over $T_{i,j}$ with (i, j) satisfying $i + j \leq n + w$, i.e.

$$\mathcal{M}_{n,w} = \max_{\substack{(i,j,\Delta)\in\mathcal{H}_n \\ i+j+2\Delta\leq n+w}} S_{i,j}^{\Delta} = \max_{\substack{(i,j) \\ i+j\leq n+w}} T_{i,j}. \tag{6.2}$$

Definition 6.2.1 *Given a window size w, a structure $(i, j, \Delta) \in \mathcal{H}_{n,w}$ with $i + j + 2\Delta \leq n + w$ is called an excursion if*

$$\begin{aligned} &T_{i,j} = 0, \quad S_{i,j}^{\delta} > 0 \text{ for } 0 < \delta < \Delta \\ &\text{and either} \quad S_{i,j}^{\Delta} = 0,\ i + \Delta = n,\ j + \Delta = n \\ &\text{or} \quad i + j + 2\Delta = n + w \end{aligned}$$

Let $\mathcal{E}_{n,w}$ be the set of all excursions with window w.

For $e = (i, j, \Delta) \in \mathcal{E}_{n,w}$ an excursion

$$\mathcal{M}_e = \max_{0<\delta\leq\Delta} S_{i,j}^{\delta} = \max_{0<\delta\leq\Delta} T_{i+\delta,j+\delta}$$

is the maximum over an excursion, and we can introduce the counting of excesses.

Definition 6.2.2 *The essentially different excesses over t is defined as*

$$C_{n,w}(t) = \sum_{e \in \mathcal{E}_{n,w}} 1(\mathcal{M}_e > t), \tag{6.3}$$

hence from (6.2) we get $(C_{n,w}(t) = 0) = (\mathcal{M}_{n,w} \leq t)$.

6.3 Stem-loops in a Markov chain

We assume that the process $(X_k)_{k\geq 1}$ is a stationary, irreducible and aperiodic Markov chain with transition probabilities P and invariant measure π. It plays a role in this chapter that the Markov chain is assumed stationary since the time reversion $(Y_k)_{1\leq k\leq n}$ is then also a time-homogeneous Markov chain. In fact, it is a stationary, irreducible and aperiodic Markov chain too with transition probabilities

$$\overleftarrow{P}_{x_0,x_1} = \frac{\pi_{x_1} P_{x_1,x_0}}{\pi_{x_0}}.$$

Assume for all $T \geq 1$ that there exists a cycle $x_1, \ldots, x_n$ (w.r.t. P) and a cycle $y_1, \ldots, y_n$ (w.r.t. to $\overleftarrow{P}$) such that

$$\sum_{k=1}^{n} f(x_k, y_k) \neq \sum_{k=1}^{n} f(x_k, y_{k+T \pmod n}).$$

Furthermore, assume that there exists a cycle $(x_1, \ldots, x_n)$ (w.r.t. P) and a cycle $(y_1, \ldots, y_n)$ (w.r.t. to $\overleftarrow{P}$) such that

$$\sum_{k=1}^{n} f(x_k, y_k) > 0.$$

Considering $(X_k)_{k\geq 1}$ as part of a doubly infinite sequence $(X_k)_{k\in\mathbb{Z}}$, we observe that

$$(Y_k)_{k=1}^{n} \stackrel{\mathcal{D}}{=} (X_k)_{k=-1}^{-n}.$$

The process $(S_n)_{n\geq 1}$ defined by $S_n = \sum_{k=1}^{n} f(X_k, X_{-k})$ is a MAP with

$$H_{(x_0,x_1),(y_0,y_1)} = \delta_{f(x_1,y_1)}$$

and transition probabilities for the underlying Markov chain being $P \otimes \overleftarrow{P}$. This Markov chain has invariant probability measure $\pi_P \otimes \pi_P$. Assuming that the invariant mean is negative,

$$\mu = \pi_P \otimes \pi_P(f) = \sum_{x,y\in E} f(x,y)\pi_P(x)\pi_P(y) < 0,$$

and defining the $E^2 \times E^2$ matrix $\Phi(\theta)$ by

$$\Phi(\theta)_{(x_0,y_0),(x_1,y_1)} = \exp(\theta f(x_1, y_1)) P_{x_0,x_1} \overleftarrow{P}_{y_0,y_1}$$

with spectral radius $\varphi(\theta) = \mathrm{spr}(\Phi(\theta))$, there is a unique solution $\theta^* > 0$ to $\varphi(\theta) = 1$. Likewise, K^* is the constant defined for this MAP by (2.19) in Theorem 2.4.3.

In addition to Φ_1 and Φ_2 defined in Section 5.3 (with $Q = \overleftarrow{P}$), we need to introduce two further Φ-matrices. Define for $\theta \in \mathbb{R}$

$$\begin{aligned}
\Phi_1(\theta)_{(x_0,y_0,z_0),(x_1,y_1,z_1)} &= \exp(\theta f(x_1, y_1) + \theta f(x_1, z_1)) P_{x_0,x_1} \overleftarrow{P}_{y_0,y_1} \overleftarrow{P}_{z_0,z_1} \\
\Phi_2(\theta)_{(x_0,w_0,y_0),(x_1,w_1,y_1)} &= \exp(\theta f(x_1, y_1) + \theta f(w_1, y_1)) P_{x_0,x_1} P_{w_0,w_1} \overleftarrow{P}_{y_0,y_1} \\
\Phi_3(\theta)_{(x_0,y_0,z_0),(x_1,y_1,z_1)} &= \exp(\theta f(x_1, y_1) + \theta f(z_1, x_1)) P_{x_0,x_1} \overleftarrow{P}_{y_0,y_1} \overleftarrow{P}_{z_0,z_1} \\
\Phi_4(\theta)_{(x_0,w_0,y_0),(x_1,w_1,y_1)} &= \exp(\theta f(x_1, y_1) + \theta f(y_1, w_1)) P_{x_0,x_1} P_{w_0,w_1} \overleftarrow{P}_{y_0,y_1}.
\end{aligned}$$

Note the subtle difference between Φ_1 and Φ_3 and between Φ_2 and Φ_4 respectively. The change is a permutation of the arguments to the second f, hence if f were symmetric there would be no difference at all. We don't want to assume that f is symmetric, and we need to take all four Φ-matrices into account. Let $\varphi_i(\theta) = \mathrm{spr}(\Phi_i(\theta))$ for $i = 1, 2, 3, 4$ denote the spectral radii of the Φ-matrices.

Theorem 6.3.1 *Assume that $\mu < 0$ and that θ^* and K^* are chosen as described above. Assume, furthermore, that*

$$\varphi_i\left(\frac{3}{4}\theta^*\right) < 1 \tag{6.4}$$

for $i = 1, 2, 3, 4$. Then if we for $x \in \mathbb{R}$ define

$$t_{n,w} = \frac{\log K^* + \log(n(w-1) - w(w-1)/2) + x}{\theta^*} \tag{6.5}$$

and $x_{n,w} \in [0, \theta^*)$ by $x_{n,w} = \theta^*(t_{n,w} - \lfloor t_{n,w} \rfloor)$, it holds that

$$||\mathcal{D}(C_{n,w}(t_{n,w})) - \mathrm{Poi}(\exp(-x + x_{n,w}))|| \to 0 \tag{6.6}$$

for $n, w \to \infty$ such that $(\log n^2)^3 = o(w)$. In particular

$$\mathbb{P}(\mathcal{M}_{n,w} \leq t_{n,w}) - \exp(-\exp(-x + x_{n,w})) \to 0 \tag{6.7}$$

for $n, w \to \infty$.

Remark 6.3.2 *The number of entries in the score matrix fulfilling that* $n+1 < i+j \leq n+w$ *can be shown to equal* $n(w-1) - w(w-1)/2$. *If* $w = n$ *we have that* $n(w-1) - w(w-1)/2 = n(n-1)/2$ *corresponding to (approximately) half of the score matrix is being used. If we fix* w *and let* $n \to \infty$ *we find, on the other hand, that* $n(w-1) - w(w-1)/2 \sim n(w-1)$. *The condition imposed on the simultaneous growth to* ∞ *of* n *and* w *may seem a little strange. It basically means that* w *can not grow arbitrarily slowly compared with* n, *and in the line of proof presented here, it turns out that* $(\log n^2)^3/w \to 0$ *is a sufficient condition. It may very well be improved slightly, but it seems irrelevant to do so.*

6.4 Proofs

We define the index set I_0 as follows

$$I_0 = \left\{ (k, ql) \mid k \in \{0, \ldots, n\},\ q \in \{0, \ldots, \left\lfloor \frac{n}{l} \right\rfloor\},\ n+1+3l \leq k+ql \leq n+w \right\}.$$

The index set I_0 is non-empty if $1+3l \leq w$ and one should generally think of $w \gg l$. Below we choose $l \sim (\log n^2)^3$, so if $(\log n^2)^3 = o(w)$ for $n \to \infty$ the set I_0 is indeed non-empty for large n. As for alignments we introduce for $t > 0$ and $a = (k, r) \in I_0$ the stochastic variable

$$V_a = V_a(t) = 1\left(\max_{1 \leq \delta \leq \Delta \leq l} \sum_{h=\delta}^{\Delta} f(X_{k+h}, Y_{r+h}) > t \right),$$

which count if there is an excursion exceeding t within the diagonal-within-a-strip indexed by a. There is a very subtle reason for restricting our attention to indexes $a = (k, r)$ satisfying $n+1+3l \leq k+r$ and not just $n+1 \leq k+r$. It corresponds to simply ignoring a sub-diagonal strip of width $3l$ in the score matrix. A strip that is asymptotically negligible for $n \to \infty$ under the conditions that $l \sim (\log n^2)^3$ and $(\log n^2)^3 = o(w)$. The reason is to get rid of some very nasty dependencies close to the diagonal, which arise due to the folding of the sequence back onto itself.

The neighbourhood of strong dependence also needs to be redefined for this setup. For $a = (k, r)$ let

$$\begin{aligned}
B_a^1 &= \{k-l, \ldots, k+2l\} \times \{0, l, 2l, \ldots, \left\lfloor \frac{n}{l} \right\rfloor l\} \cap I_0 \\
B_a^2 &= \{0, \ldots, n\} \times \{r-l, r, r+l\} \cap I_0 \\
B_a^3 &= \{0, \ldots, n\} \times \{n-k+1-l, n-k+1, n-k+1+l\} \cap I_0 \\
B_a^4 &= \{n-r+1-l, \ldots, n-r+1+2l\} \times \{0, l, 2l, \ldots, \left\lfloor \frac{n}{l} \right\rfloor l\} \cap I_0
\end{aligned}$$

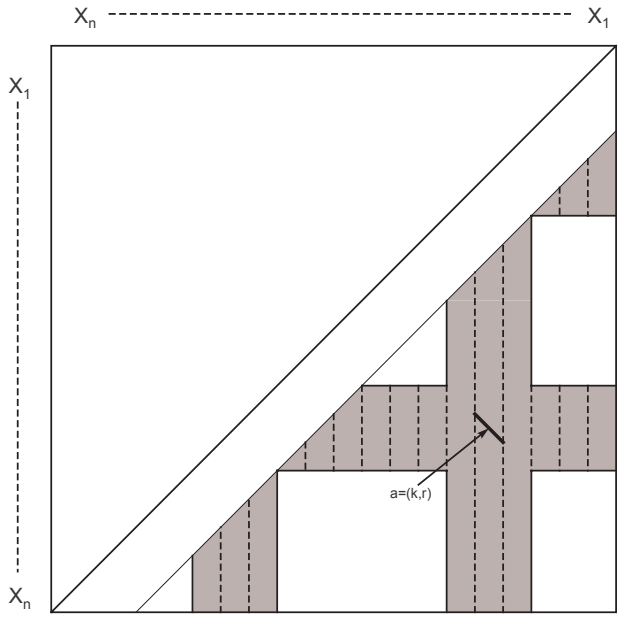

Figure 6.1: An example of the neighbourhood of strong dependence for the diagonal-within-a-strip, a, in the case of folding a sequence back onto itself. The neighbourhood becomes slightly more complicated in this case compared to aligning two independent sequences, because the reuse of variables is more intense. Here we see the case where the windows size $w = n$.

and put $B_a = B_a^1 \cup B_a^2 \cup B_a^3 \cup B_a^4$. Intersecting with I_0 in the definitions above is a restriction of the four sets to the triangular part of the score matrix we are interested in. On Figure 6.1 we see an example of the neighbourhood of strong dependence in the case where $w = n$.

Lemma 6.4.1 *If we, for some $x \in \mathbb{R}$, let*

$$l = l_n \sim (\log n^2)^3 \quad \text{and} \quad t = t_{n,w} = \frac{\log K^* + \log(n(w-1) - w(w-1)/2) + x}{\theta^*}$$

and define $x_{n,w} \in [0, \theta^*)$ by $x_{n,w} = \theta^*(t_{n,w} - \lfloor t_{n,w} \rfloor)$, then under the assumptions in Theorem 5.3.1, the conditions in Theorem 5.4.1 are fulfilled with

$$\lambda_n = \exp(-x + x_n).$$

That is

$$\left|\left| \mathcal{D}\left(\sum_{a \in I_0} V_a\right) - \text{Poi}(\exp(-x + x_{n,w})) \right|\right| \to 0.$$

Proof: The proof is to some extend a copy of the proof of Lemma 5.5.11. There are, however, some details that need to be handle carefully.

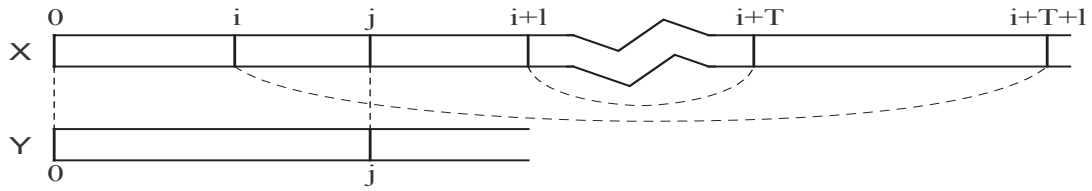

Figure 6.2: In the fold-back setup there are, in addition to the two different cases from the ordinary alignment setup, also two essentially different cases where the overlapping part is turned around before realignment. Either as shown here the overlap is in the X-sequence, or the overlap is in the Y-sequence (which is the X-sequence reversed, but there is a difference).

First of all observe that due to the assumption that $(\log n^2)^3/w \to 0$ for $n, w \to \infty$ it holds that $|I_0| \sim (n(w-1) - w(w-1)/2)l^{-1}$. And since

$$\mathbb{P}(V_a > t_n) = \mathbb{P}(V_a > t_n - x_n/\theta^*) \sim (n(w-1) - w(w-1)/2)^{-1} l \exp(-x + x_n)$$

as follows from Lemma 2.4.6 (the convergence is uniform in a), we have that

$$\sum_{a \in I_0} \mathbb{E}(V_a) \sim |I_0|(n(w-1) - w(w-1)/2)^{-1} l \exp(-x + x_n) \sim \exp(-x + x_n),$$

or, since $\exp(-x + x_n)$ is bounded, $\left|\sum_{a \in I_0} \mathbb{E}(V_a) - \exp(-x + x_n)\right| \to 0$
Furthermore, $\mathbb{E}(V_a)\mathbb{E}(V_b) \sim (n(w-1) - w(w-1)/2)^{-2} l^2 \exp(-2(x + x_n))$ and $\max_a |B_a| = O(n) = o(|I_0|)$, so

$$\sum_{a \in I, b \in B_a} \mathbb{E}(V_a)\mathbb{E}(V_b) \leq K|I_0| \max_a |B_a|(n(w-1) - w(w-1)/2)^{-2} l^2 \to 0$$

for $n, w \to \infty$.

To verify that condition (5.17) in Theorem 5.4.1 holds, we divide the neighbourhood of strong dependence B_a into five disjoint sets;

$$C_a = B_a^1 \cap B_a^2, \quad D_a^1 = B_a^1 \backslash C_a, \quad \text{and} \quad D_a^2 = B_a^2 \backslash C_a$$

together with B_a^3 and B_a^4. By exactly the same arguments as in the proof of Lemma 5.5.11, it follows that

$$\begin{aligned} \sum_{a \in I_0, b \in C_a} \mathbb{E}(V_a V_b) &\to 0 \quad \text{and} \\ \sum_{a \in I_0, b \in D_a^i} \mathbb{E}(V_a V_b) &\to 0 \end{aligned}$$

for $i = 1, 2$ using for the second limit that $\varphi_i(3/4\theta^*) < 1$ for $i = 1, 2$. To take care of $a \in I_0$ and $b \in B_a^i$ for $i = 3, 4$ we need the conclusion from Lemma 5.5.3 but for the overlap as depicted in Figure 6.2. The result is easily verified by a proof similar to that of Lemma 5.5.3. Thus due to the assumption $\varphi_i(3/4\theta^*) < 1$ for $i = 3, 4$ we can obtain, for some $\varepsilon > 0$, the estimate

$$\mathbb{E}(V_a V_b) \leq K l^4 \exp(-3/2\theta^*(1+\varepsilon)t)$$

for $a \in I_0$ and $b \in B_a^i$ for $i = 3, 4$. And therefore

$$\begin{aligned} \sum_{a \in I_0, b \in B_a^i} \mathbb{E}(V_a V_b) &\leq K l^4 (n(w-1) - w(w-1)/2)^{-3/2(1+\varepsilon)} \sum_{a \in I_0} |B_a^i| \\ &\leq K l^4 (n(w-1) - w(w-1)/2)^{-3/2(1+\varepsilon)} \\ &\quad \times K_1 (n(w-1) - w(w-1)/2)^{3/2} \to 0. \end{aligned}$$

Finally, we need to verify condition (5.18). For this, we embed $(X_k)_{k \geq 1}$ into a doubly infinite, stationary Markov chain $(X_k)_{k \in \mathbb{Z}}$, which is exponentially β-mixing, thus

$$\beta(k) = \mathbb{E} \sup_{A \in \mathcal{F}_{[k,\infty)}} |\mathbb{P}(A|\mathcal{F}_{(-\infty,0]}) - \mathbb{P}(A)|, \leq K_1 \exp(-K_2 k).$$

For $a = (k, r)$ we let $m = n - r + 1 + 3l \leq k$ and define $I_1 = (-\infty, m - l]$, $I_2 = [m+1, m+l]$, $I_3 = [m+2l+1, k-l]$, $I_4 = [k+1, k+l]$ and $I_5 = [k+2l+1, \infty)$. With $I = I_1 \cup I_3 \cup I_5$ and $J = I_2 \cup I_4$ we clearly have that $\mathcal{F}_a \subseteq \mathcal{F}_I = \sigma(X_n \mid n \in I)$ and V_a is $\mathcal{F}_J$ measurable. Hence from Lemma 2.6.2 and Theorem 2.6.4 we obtain that

$$\mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)| \leq 7\alpha(\mathcal{F}_I, \mathcal{F}_J) \leq 7\beta(l+1) \leq K \exp(-K_2 l).$$

This implies that

$$\sum_{a \in I_0} \mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)| \leq K(n(w-1) - w(w-1)/2)^2 \exp(-K_2(\log n^2)^3) \to 0$$

for $n, w \to \infty$. □

Proof of Theorem 6.3.1 The idea is again to show that

$$\mathbb{P}\left(\sum_{a \in I_0} V_a(t_{n,w}) \neq C(t_{n,w})\right) \to 0,$$

and the technique is the same as in the proof of Theorem 5.3.1. That is, we can show that all the excursions in $\mathcal{E}_{n,w}$ exceeding t occur with probability tending to

1 in such a way that they each and every one of them are counted exactly once by $\sum_{a \in I_0} V_a(t_{n,w})$.

The only difference where one has to be a little careful is what happens close to the boundary and with the sub-diagonal strip that we ignored in I_0. The probability of any excursion exceeding t occurring in these regions can, however, as in the proof of Theorem 5.3.1 be seen to be bounded by $Knl \exp(-\theta^* t) \to 0$ for $n, w \to \infty$ – using again that $l/w = (\log n^2)^3/w \to 0$. □

6.5 Discussion

Though the result is very similar to that of Chapter 5, it is a genuinely new result – even if we consider a sequence of *iid* variables – and it seems to be a novel idea to take this approach to search for local structures in sequences. It should be emphasised that we look specifically for local stem-loop structures, and this is what makes the theory very similar to that of local alignments.

If the reader is confused by the introduction of a window, it may be beneficial to first think of $w = n$, which corresponds to no window. Introducing a window effectively means to restrict our attention to a part of the score matrix. It is then no surprise that the Poisson approximation is still valid with a mean that is suitably modified.

7

Searching for Small Folding Structures

7.1 Introduction

The development of the results in Chapter 6 was inspired by some recent discoveries in biology. Certain small RNA-molecules were found to play a role in the regulation of gene expression. It was observed that RNA can interfere with the translation of messenger RNA (mRNA) by complementary binding to the mRNA and thereby preventing translation or directing degradation of the mRNA (C. Lee & Ambros 2001, Lau et al. 2001, Lagos-Quintana et al. 2001). It was also found that RNA-molecules with this function were transcribed from the cells own DNA, and such RNA-molecules were, due to their size of approximately 21-22 nucleotides, named micro RNA (miRNA). A characteristic feature of miRNA is that the molecule is transcribed into a so-called *precurser* of size approximately 100 nucleic acids possessing a stem-loop structure. This structure seems to be important for the way the molecule is afterwards 'prepared' by cutting out the actual miRNA from the precurser.

Our interest in miRNA is as an example of RNA-genes or non-coding, functional RNA. By RNA-genes we mean chromosomal DNA that is transcribed into RNA and show some function in the cell other than 'just' being messenger RNA. A classical example is transfer RNA (tRNA) that assists the ribosome (which also contains RNA-parts) in the translation of mRNA into protein. We are interested in the finding of such RNA-genes in the DNA by computational methods, and the main idea presented here is to search for small sequence parts that posses a stem-loop structure. The stem-loop is a characteristic feature for miRNA that seems to be appropriate for

distinguishing miRNA, but for other kinds of RNA-genes other (structural) features might be more important.

The main points treated in this chapter are how the miRNA stem-loop structure can be modelled and used to find miRNA-like RNA in large DNA-sequences, and how we can evaluate, improve and optimise the search technique based on the theory developed in Chapter 6 (and some extrapolation). In this setup we also discuss, using simulations, the applicability of the extreme value theory developed.

7.2 Modelling miRNA

We present a dataset consisting of 59 miRNAs from the worm *C. elegans* for which we fit relevant models. Null-models are fitted using the *C. elegans* genome, and we construct score functions by the approach in Chapter 4. The dataset contains 59 sequences from the alphabet

$$E = \{\text{A}, \text{C}, \text{G}, \text{U}\}$$

each consisting of between 72 and 110 letters. This dataset was kindly extracted in 2002 by Morten Lindow, University of Copenhagen, from the Rfam database (Griffiths-Jones et al. 2003). Today, the number of miRNAs found in *C. elegans* has increased in the database, but we strick to the original dataset for this analysis. Together with the sequence of letters comes a *predicted* structure, which is found by computational methods. Using predicted structures for the estimation of a model may seem as a problem, but we argue that this is not so. The point is that we want to fit a model for the sequence of letters for a given structure, which we believe to be a stem-loop. The structure prediction programs predict structure from the sequence as the minimal energy structure under a specific choice of energy function – or one may prefer to say that the predicted structure is the most likely such structure under a corresponding Gibbs measure. In all cases the predicted structure turns out to be a stem-loop for the dataset considered. Hopefully, this transformation of the sequence, which we use to fit a model, captures some of the essential features of the letter composition in the miRNA. We can view the use of predicted structures as a missing data problem, which we solve by imputation. Though this may not be optimal, we will not pursue a discussion of alternative methods.

In Figure 7.1 we see an example of a miRNA from *C. elegans* called mir-1. The figure shows the characteristic stem-loop structure of the precurser for miRNA. We see the hairpin loop at the right and a few internal loops and bulges along the stem. Most of the letters form canonical Watson-Crick pairs, that is AU and CG-pairs.

```
5′−GUG      UA     AG                C        GC              AUC
       ACCG   CCG   C UGCAUACUUC  UUACAU   CCA UA  CUAU     A
       UGGU   GGC   G[AUGUAUGAAG  AAUGUA   GGU]AU  GGUA     U
3′−UGA      GG     AA                A        A         A      AA
```

Figure 7.1: The miRNA mir-1 ($5' -$ UGGAAUGUAAAGAAGUAUGUA $- 3'$ in square brackets) in the surrounding precurser, which form a (predicted) stem-loop structure.

Together with GU pairs, which also occur a few times in the structure, we refer from hereon to such pairs as canonical pairs. Pairs different form the canonical pairs are not found in the structure. For instance, around the middle of the stem we find a lonely C on the upper strand opposing a lonely A on the lower, but they are not regarded as forming a pair - on the contrary, they form an small internal loop. If we only allow for canonical pairs we end up with a model and a score function that prohibits all other pairs in a structure. Thus the stem structure for the mir-1 precurser would essentially be broken up into six contiguous parts formed by allowed pairs only. Without a finite gap penalty function for the internal loops, the score function would not be able to 'connect' these six parts together. This is of course undesirable. We could just allow for a finite gap penalty function, but we want to make an analysis first without introducing gaps to be able to apply the theory developed. A solution is to regard the *symmetric* internal loops as if they also form pairs. We therefore basically include all the symmetric loops, which pari up in an obvious way, into the structure and then fit a model to the data. The internal loops are short and rather rare compared to the number of canonical pairs, but with the inclusion of these non-canonical pairs we obtain a non-degenerate model and thus a score function allowing for all pairs to occur. Moreover, since the non-canonical pairs are rather unlikely, they receive a negative score as we show below.

With this approach of preparing the data we obtain, using maximum likelihood estimation, the estimate $\hat{\nu}$ specifying the alternative hypotheses under the *iid* model. Using chromosome 1 (one of the strands) from *C. elegans* we obtain an estimate $\hat{\lambda}$ specifying the null hypothesis. There are 2220 pairs in the miRNA dataset used to fit ν and roughly 15 million letters in chromosome 1 used to fit λ.

	$\hat{\nu}$				$\hat{\lambda}$
	A	C	G	U	
A	0.01	0.01	0.01	0.20	0.32
C	0.01	0.01	0.21	0.01	0.18
G	0.01	0.18	0.01	0.05	0.18
U	0.21	0.01	0.05	0.01	0.32

Observe that had we used the complementary strand in chromosome 1 for the estimation of λ, we would for these data obtain the same estimate. These estimates result in a log-likelihood ratio score function f being:

	A	C	G	U
A	−2.30	−1.59	−2.07	0.64
C	−1.35	−1.62	1.89	−2.08
G	−1.71	1.74	−1.55	−0.19
U	0.70	−1.63	−0.10	−1.94

Note the some of the asymmetry appearing in f compared to $\hat{\nu}$ is a result of using a higher precision of $\hat{\nu}$ in the computation of f than the two digits given above. As discussed in Chapter 1, we typically use an integer version of the score function, thus consider[1] instead of f the score function $f_1 = I_4(f)$:

	A	C	G	U
A	−9	−6	−8	3
C	−5	−6	8	−8
G	−7	7	−6	−1
U	3	−7	0	−8

We can make a few interesting observations about this score function. First of all, the positive entries are not located on the diagonal but instead (with the ordering of the letters used here) on the 'anti'-diagonal. This is different from the score functions used for alignment, and of course it just reflects that the preferred pairs in a structure are pairs between different letters (canonical pairs) whereas preferred pairs in an alignment are pairs between equal letters. Secondly, the score function is derived from $\hat{\nu}$ with marginals clearly differing from $\hat{\lambda}$. Thus the score function actually detects a combination of potential structural pairing and letter composition bias. This results for instance in that CG pairs score more than twice as much as AU pairs even though the two kinds of pairs are almost equally likely under the $\hat{\nu}$-measure. This phenomenon is sometimes referred to as a CG-bias, but perhaps it is more correct to say that it is the genome that shows an AU-bias. It is of relevance to investigate whether the score function just detects CG-bias or whether there is more to it. We continue this discussion in the next section.

Furthermore, one can observe that the score function is slightly asymmetric, but this author is not convinced that this is really important. Asymmetry can, however, not be ruled out as a reasonable phenomenon. The asymmetry of the CG- versus the GC-pair estimated for $\hat{\nu}$ corresponds to a (small) preference of the C to be in the $3'$-strand of the stem-loop.

[1] we choose to keep the score function of moderate size to be able to apply the computational methods discussed in Chapter 2 for computing K^*, hence the moderate multiplication by 4

A corresponding estimation is carried out using the Markov model. A problem occurs due to the limited dataset and the many parameters in the transition matrix R specifying the alternative hypotheses under the Markov model. Roughly half of the transitions between non-canonical pairs don't occur in the dataset even though we include the symmetric internal loops as pairs in the structures. This is undesirable as it will again lead to a score function that prohibits certain transitions. We make an *ad hoc* solution and add a pseudo-count to all entries in the matrix of transition counts, that is, we simply add one to all the entries. This results in the matrix, $\hat{R}$, of estimated transition probabilities based on 1972 transitions:

	AA	CA	GA	UA	AC	CC	GC	UC	AG	CG	GG	UG	AU	CU	GU	UU
AA	0.05	0.08	0.05	0.10	0.03	0.05	0.15	0.03	0.05	0.15	0.03	0.03	0.10	0.03	0.05	0.03
CA	0.02	0.04	0.02	0.10	0.06	0.02	0.12	0.04	0.04	0.18	0.02	0.10	0.12	0.02	0.06	0.02
GA	0.05	0.05	0.05	0.10	0.03	0.03	0.10	0.05	0.03	0.18	0.03	0.05	0.13	0.03	0.03	0.08
UA	0.01	0.02	0.01	0.21	0.00	0.01	0.19	0.00	0.00	0.20	0.02	0.04	0.19	0.02	0.05	0.01
AC	0.02	0.02	0.02	0.14	0.02	0.02	0.19	0.02	0.02	0.12	0.05	0.02	0.19	0.02	0.05	0.05
CC	0.03	0.03	0.03	0.20	0.03	0.03	0.13	0.03	0.03	0.17	0.03	0.03	0.10	0.03	0.03	0.03
GC	0.01	0.02	0.01	0.27	0.02	0.01	0.14	0.01	0.01	0.16	0.00	0.07	0.19	0.01	0.05	0.02
UC	0.02	0.05	0.02	0.14	0.02	0.02	0.12	0.02	0.02	0.19	0.02	0.02	0.19	0.02	0.05	0.05
AG	0.06	0.06	0.12	0.06	0.09	0.03	0.06	0.03	0.03	0.09	0.03	0.09	0.12	0.03	0.03	0.03
CG	0.01	0.01	0.02	0.22	0.02	0.01	0.15	0.03	0.01	0.21	0.01	0.03	0.20	0.01	0.05	0.02
GG	0.03	0.03	0.03	0.19	0.03	0.03	0.13	0.03	0.03	0.06	0.03	0.10	0.13	0.03	0.06	0.03
UG	0.01	0.02	0.01	0.19	0.03	0.03	0.18	0.03	0.01	0.20	0.02	0.04	0.17	0.02	0.04	0.02
AU	0.02	0.01	0.01	0.12	0.01	0.01	0.26	0.01	0.01	0.21	0.01	0.06	0.19	0.01	0.05	0.01
CU	0.03	0.03	0.03	0.06	0.03	0.03	0.22	0.03	0.03	0.19	0.03	0.03	0.09	0.03	0.03	0.09
GU	0.01	0.04	0.02	0.17	0.02	0.02	0.15	0.03	0.01	0.25	0.02	0.09	0.08	0.02	0.06	0.03
UU	0.02	0.04	0.02	0.16	0.06	0.04	0.10	0.06	0.06	0.12	0.02	0.04	0.14	0.02	0.02	0.06

The transition probabilities P (and $\overleftarrow{\hat{P}}$) under the null hypothesis are estimated again using chromosome 1 from *C. elegans* with roughly 15 million transitions:

$\hat{P}$

	A	C	G	U
A	0.42	0.15	0.16	0.27
C	0.34	0.19	0.19	0.28
G	0.35	0.19	0.19	0.26
U	0.19	0.20	0.19	0.42

$\overleftarrow{\hat{P}}$

	A	C	G	U
A	0.42	0.19	0.20	0.19
C	0.26	0.19	0.19	0.35
G	0.28	0.19	0.19	0.34
U	0.27	0.16	0.15	0.42

Observe that for the permutation $\sigma = (\text{AU})(\text{CG})$ we apparently have that

$$\overleftarrow{\hat{P}}_{\sigma(x),\sigma(y)} = \hat{P}_{x,y}.$$

There is no particular reason that this should be the case since, in general, for an arbitrary P

$$\overleftarrow{P}_{\sigma(x),\sigma(y)} = \frac{\pi_{\sigma(y)} P_{\sigma(y),\sigma(x)}}{\pi_{\sigma(x)}}.$$

The transition probabilities given by $\overleftarrow{\hat{P}}_{\sigma(x),\sigma(y)}$ are the transition probabilities for the complementary strand of the strand we used for estimating $\hat{P}$. It is interesting, like for the *iid* case, that there is no apparent difference between the Markov chains modelling the two strands.

The corresponding integer version, $f_2 = I_4(f)$, of the score function becomes:

	AA	CA	GA	UA	AC	CC	GC	UC	AG	CG	GG	UG	AU	CU	GU	UU
AA	−5	0	−2	1	−4	3	7	0	−1	7	−1	−1	0	−3	0	−3
CA	−7	−3	−6	−2	2	−1	6	−1	0	7	−2	2	2	−4	1	−6
GA	−3	−2	−2	−1	−2	0	5	0	−2	7	−1	0	2	−3	−3	−1
UA	−9	−4	−7	1	−9	−5	9	−10	−9	8	−1	−2	4	−4	1	−9
AC	−7	−4	−4	3	−5	−2	6	−2	−5	5	1	−2	2	−3	−1	0
CC	−4	−3	−3	2	−2	0	5	−3	−2	6	0	−3	1	−2	−2	−4
GC	−9	−5	−7	3	−4	−6	5	−6	−6	6	−10	0	4	−6	0	−6
UC	−5	0	−3	0	−3	−1	6	−5	−3	7	−1	−5	4	−2	1	−4
AG	−4	0	2	0	1	−1	2	−1	−4	4	−1	4	0	−2	−2	−2
CG	−7	−8	−5	2	−4	−6	6	−3	−6	7	−6	−3	4	−8	0	−7
GG	−5	−3	−3	2	−2	0	5	−3	−2	2	−1	2	2	−2	1	−4
UG	−10	−5	−7	1	−3	0	7	−5	−7	8	−2	−3	3	−4	0	−8
AU	−5	−4	−4	5	−10	−6	8	−7	−7	7	−4	2	0	−8	−2	−7
CU	−2	−1	−1	0	−2	−1	7	−3	−2	7	−1	−3	−1	−4	−4	−2
GU	−7	0	−3	4	−4	−3	5	−4	−7	8	−3	1	−2	−6	−1	−7
UU	−4	1	−1	3	1	1	5	−1	1	6	−1	−3	1	−5	−4	−4

It is hard to make an intelligent summary of the content of these 256 numbers. The Watson-Crick canonical pairs have been framed to make the reading of the table easier – it is after all most important how these pairs score. Scrutinising the table might reveal interesting observations. For instance, the 'alternating' transition UA to AU scores higher than UA to UA. This can be explained by observing that transitions under the null hypothesis from A to A are twice as likely as transitions from A to U. And likewise for U to U transitions. On the other hand, under the alternative hypothesis, the alternating transition UA to AU are almost as likely as the non-alternating transition UA to UA, which makes 'alternation' a distinguishing feature of the letter composition in miRNA. Otherwise, one should observe that transitions to the canonical pairs in general receive a positive score whereas transitions to non-canonical pairs receive a negative score. Moreover, the CG-bias phenomenon can also be observed for this score function.

7.3 Detecting miRNA

With the score functions found in the preceding section we can try to detect miRNA-precursers within long sequences of DNA. For a given sequence $\mathbf{x}$ from E of length n the method is simply to construct the score matrix $T = (T_{ij})_{0 \le i,j \le n}$ given by (6.1) – using $\mathbf{x}$ as a realisation of the stochastic process $(X_n)_{n \ge 1}$ – and then search for high scoring excursions. If we fix the window size w prior to computing T, we can save some computations. Realistic choices of n and w could be $n = 15 \times 10^6$ and $w = 200$. This would roughly correspond to searching a single *C. elegans* chromosome using a window of size 200. We choose 200 as a window size sufficiently large for capturing all the miRNAs in the dataset.

Then fixing a threshold s_0, cf. the discussion in Section 4.4, we can classify the excursions with a maximum exceeding s_0 as stem-loop structures and hence as potential miRNAs. We want to use Theorem 6.3.1, given that the conditions imposed for the theorem to hold are actually fulfilled, to approximate the specificity

$$1 - \alpha(s_0) = \mathbb{P}_{\lambda_0}\left(\mathcal{M}_{n,w} \le s_0\right).$$

It follows from Theorem 6.3.1 that the distribution of

$$\mathcal{M}^*_{n,w} = \theta^* \mathcal{M}_{n,w} - \log K^*$$

asymptotically doesn't depend on λ_0 nor on the score function f. This is a highly useful observation and this normalisation of the maximal score will serve to make the performance of different score functions comparable. With

$$\alpha^*(s_0) = \alpha((s_0 + \log K^*)/\theta^*),$$

Theorem 6.3.1 implies that

$$1 - \alpha^*(s_0) \simeq \exp(-(n(w-1) - w(w-1)/2)\exp(-s_0 + s_{n,w})) \tag{7.1}$$

with $s_{n,w} \in [0, \theta^*)$. Furthermore, the distribution of $C^*_{n,w}(s_0) = C_{n,w}((s_0 + \log K^*)/\theta^*)$, the number of normalised excursion maxima exceeding s_0, can be approximated by a Poisson distribution with mean

$$\xi_{n,w}(s_0) = (n(w-1) - w(w-1)/2)\exp(-s_0 + s_{n,w}). \tag{7.2}$$

Fixing w, it follows from (7.2) that if $\xi_{n,w}(s_0)$ is to be kept fixed, the threshold scales approximately like $\log n$. That is to say, changing n we should change the threshold s_0 roughly like $\log n$ to keep the mean $\xi_{n,w}(s_0)$ fixed. The asymptotic approximations are of course in principle only valid for n and w tending to infinity and $\xi_{n,w}(s_0)$ having

a non-zero limit. Nevertheless, the approximation is often quite good for reasonable choices of n, w and s_0 of practical interest, cf. Section 7.4 below.

Having established the approximation (7.1) we have an idea about how the specificity depends on the threshold. We will use the dataset to establish an empirical approximation of the sensitivity as discussed in Section 4.4. Choosing a score function and a null model we can, for each of the 59 sequences in the miRNA dataset, compute the score matrix and the corresponding *normalised* maximal score. This result in a set of 59 normalised scores, $s_1^*, \ldots, s_{59}^*$, and a corresponding empirical sensitivity

$$\beta_\varepsilon^*(s_0) = \frac{1}{59} \sum_{k=1}^{59} 1(s_k^* > s_0).$$

Thus $\beta_\varepsilon^*(s_0)$ is the fraction of miRNAs in the dataset that would be found using the threshold s_0. It is of interest to understand the covariation of the specificity and sensitivity as we change s_0. In this setup, the (asymptotic) specificity is a monotonely increasing function of s_0 (depending on n and w), and the function $s_0 \mapsto \beta_\varepsilon^*(s_0)$ therefore contains all the information we need. Moreover, it is independent of n and w and due to the normalisation it is also independent of the model parameters as well as the score function. The behaviour of the function $s_0 \mapsto \beta_\varepsilon^*(s_0)$ is therefore suitable for comparing different models and different score functions.

Assume, for example, that we want to search for potential miRNAs using the independence model. That is, we choose the score function f_1, and we need to compute the θ^* and K^* parameters under the relevant null hypothesis. Under the independence model we assume that the sequence is a sequence of *iid* variables with distribution $\hat{\lambda}$. The average score using f_1 under $\hat{\lambda}$ is found to be $\mu = -3.45$. We may, however, also be interested in other null hypothesis, e.g. a Markov chain or another *iid* model, to investigate the performance of the score function f_1 in more details. First of all, if we believe that the Markov chain given by $\hat{P}$ is a better description of random sequences than the *iid* model, it might be more appropriate to evaluate the performance of f_1 using the Markov chain as null hypothesis. Under the Markov chain null hypothesis the average score is still -3.45 due to the fact that the invariant measure of $\hat{P}$ is $\hat{\lambda}$. Secondly, as discussed in the preceding section, it is also of interest to investigate whether the score function f_1 simply detects CG-bias. To do so we consider the, relatively to $\hat{\lambda}$, CG-rich uniform distribution $\lambda_u = (0.25, 0.25, 0.25, 0.25)$ (which is also close to the marginals of $\hat{\nu}$). This results in an average score being $\mu_u = -3.13$ compared with the average -3.45 under $\hat{\lambda}$. The fact that the average is still negative and rather close to the average under $\hat{\lambda}$ show that the score function detects more than just CG-bias. We compute in these three cases the following values of the constants θ^* and K^*:

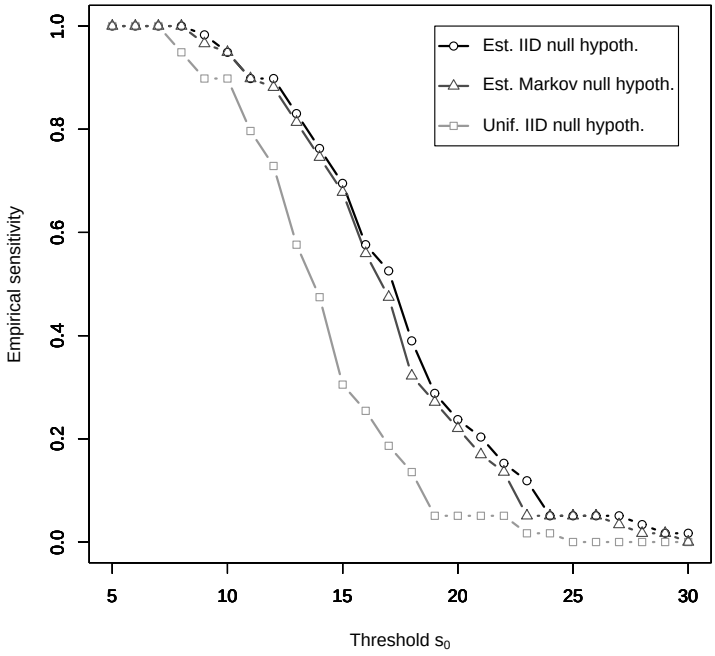

Figure 7.2: The empirical sensitivity as a function of the normalised threshold gives, for the score function f_1, a comparable picture of the performance under various null hypothesis. Here we compare the *iid* null hypothesis $\hat{\lambda}$ with the Markov chain null hypothesis given by $\hat{P}$ and the *iid* null hypothesis λ_u corresponding to a CG-bias.

	θ^*	K^*
$\hat{\lambda}$	0.232	0.201
$\hat{P}$	0.224	0.193
λ_u	0.182	0.179

One also verifies that the $\varphi_i(3/4\theta^*) < 1$ condition for $i = 1, 2, 3, 4$ is fulfilled in all three cases, and it is trivial to check that the rest of the conditions for Theorem 6.3.1 to hold are fulfilled too.

Computing the empirical scores of the miRNAs using f_1 and normalising using either of the null hypotheses give, as shown in Figure 7.2, three different empirical sensitivity functions. Due to the normalisation they are comparable. We observe that the difference between the *iid* null hypothesis $\hat{\lambda}$ and the Markov null hypothesis $\hat{P}$ is negligible. The picture is a little different when comparing with the uniform null hypothesis. The difference between the two empirical sensitivity functions, comparing the two *iid* hypothesis $\hat{\lambda}$ and λ_u, quantifies in some sense the information contained in the score function due to the CG-bias in miRNA. For example, assume that we want to search a sequence of length 100,000 using a window of size 200, and that we want the sensitivity to be approximately 80%. On Figure 7.2 we read of that to reach the desired sensitivity the normalised threshold s_0 should be approximately

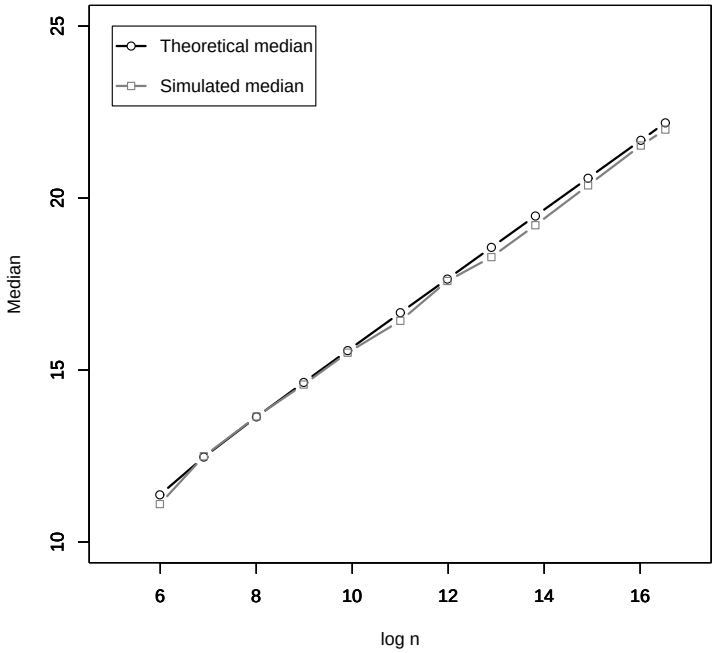

Figure 7.3: The asymptotic extreme value theory predicts that the (normalised) maximal local structure score should grow roughly like $\log n$ for fixed w. This figure shows under the *iid* model the median of the simulated maximal score as a function of $\log n$ for $w = 200$ and n ranging from 400 to 15 millions together with the asymptotic theoretical median.

13 under $\hat{\lambda}$ and 11 under λ_u. By (7.2) we find that under the CG-rich null hypothesis λ_u there will be roughly $\exp(2) \simeq 7.4$ times as many randomly occurring stem-loops exceeding the threshold as under $\hat{\lambda}$. In numbers, fixing the sensitivity to be 80% we have that $\xi_{n,w}(s_0) \simeq 45$ under $\hat{\lambda}$ and $\xi_{n,w}(s_0) \simeq 334$ under λ_u. Thus in a CG-rich sequence (as modelled by λ_u) the score function f_1 clearly looses some of it ability to distinguish true stem-loops from random stem-loops compared to an 'average' sequence as modelled by $\hat{\lambda}$. The difference is, however, not alarming, and the CG-bias is definitely not the only feature of miRNAs that f_1 captures.

7.4 A simulation study

To test the validity of the extreme value theory we conducted a simulation study. For relevant applications, as discussed above, the window size w is usually much smaller than the length n of the sequence in which we search for local stem-loop structures. We fix a rather small window size $w = 200$ and compare the theoretical results with simulations for n ranging from 400 to 15 millions. We use the score functions computed above and carry out two studies – one where we simulate under the *iid* model and use the corresponding *iid* score function and one where we simulate

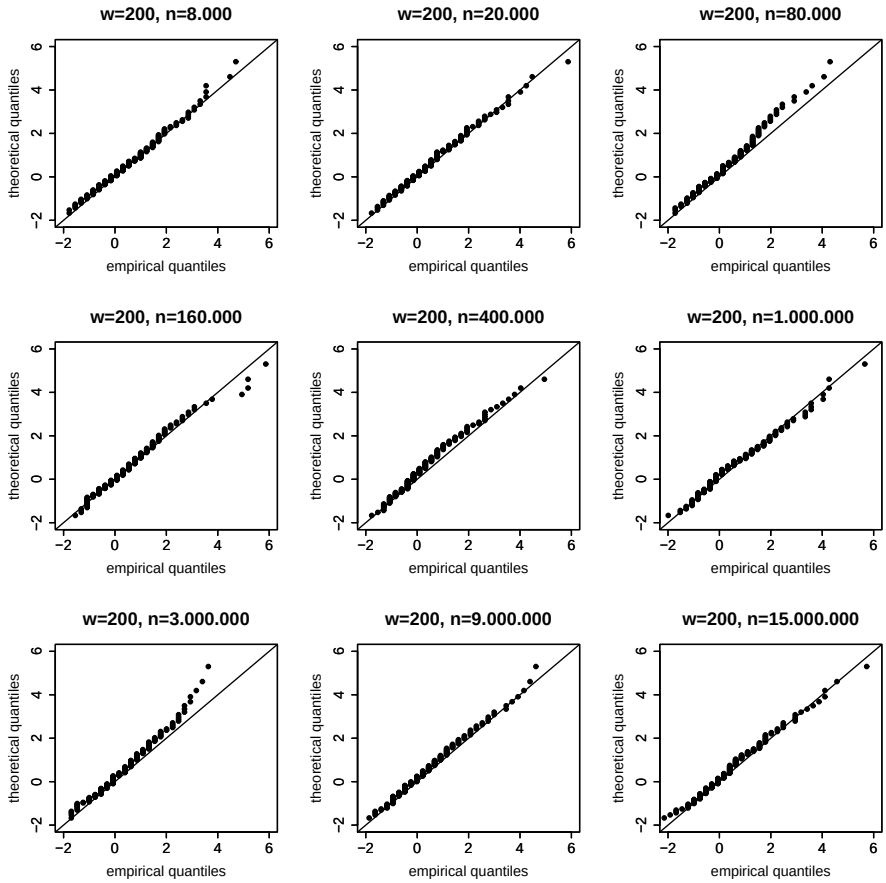

Figure 7.4: By the asymptotic extreme value theory, the normalised maximal local structure score should follow a Gumbel distribution. The figure shows under the *iid* model the QQ-plots of the normalised empirical scores versus the theoretical Gumbel distribution for $w = 200$ and n ranging from 8000 to 15 millions.

under the Markov model and use the corresponding Markov score function. In both simulation studies we simulate, for each n, 200 sequences of length n under the null hypothesis and compute the maximal local structure score using the relevant score function. As discussed, the normalised score

$$\mathcal{M}^*_{n,w} = \theta^* \mathcal{M}_{n,w} - \log K^*$$

has an asymptotically parameter independent distribution. We compare the normalised simulated scores with the relevant Gumbel distribution. We investigate the following two issues; (i) does the normalised score grow roughly like $\log n$ as indicated by the theory – even if w is fixed and much smaller than n, and (ii) does the asymptotic Gumbel distribution fit the simulated scores?

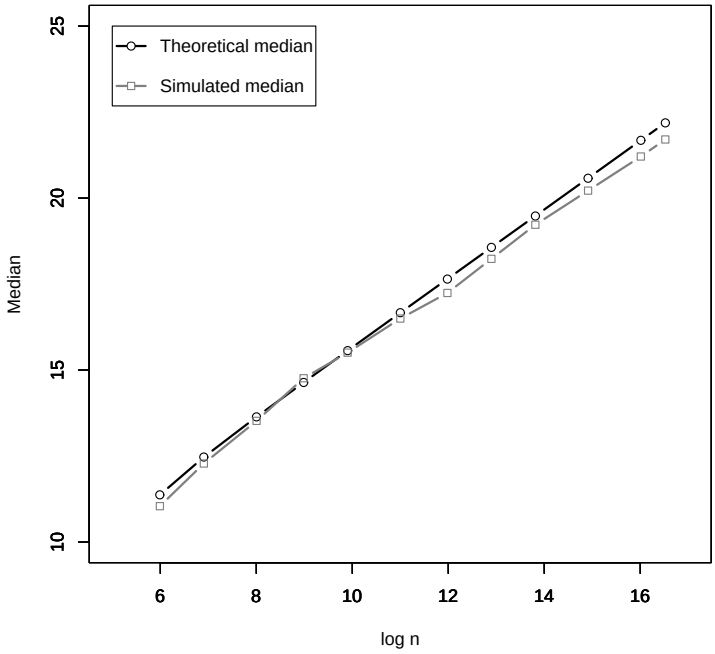

Figure 7.5: This figure shows under the Markov model the median of the simulated maximal score as a function of $\log n$ for $w = 200$ and n ranging from 400 to 15 millions together with the asymptotic theoretical median.

We have summarised the results of the *iid* simulation study in Figure 7.3 and Figure 7.4, which show the normalised empirical and theoretical medians as a function of $\log n$ and the QQ-plots of the empirical versus the theoretical quantiles. The empirical scores are in a surprisingly good agreement with the theoretical results. On Figure 7.3 we see that the median of the maximal score indeed seems to grow roughly like $\log n$ and that it falls close to the theoretical median in the whole range. There is, however, a slight tendency that the theory overestimates the median for large n. This is not surprising and can be explained as an effect of the fixed window size. As $n \to \infty$ and $w = 200$ is fixed, the size of the local structure that is needed to reach the theoretically predicted maximal local score comes in conflict with the limitations imposed by the window. In fact, the size of the local structures that can be found is bounded due to the fixed window size and there is an upper limit on the size of the maximal score that can be achieved. As $n \to \infty$ and w is fixed the maximal local structure score will thus converge to a fixed finite value almost surely. Despite of this, the theory still fits well in the wide range of n's considered here. Moreover, the QQ-plots clearly show that the Gumbel distribution is an adequate approximation of the fluctuations of the maximal score. There is a lattice effect which can be observed on the QQ-plot as a vertical clumping of the empirical scores. Despite of the lattice effect, the Gumbel distribution seems to fit the empirical scores well.

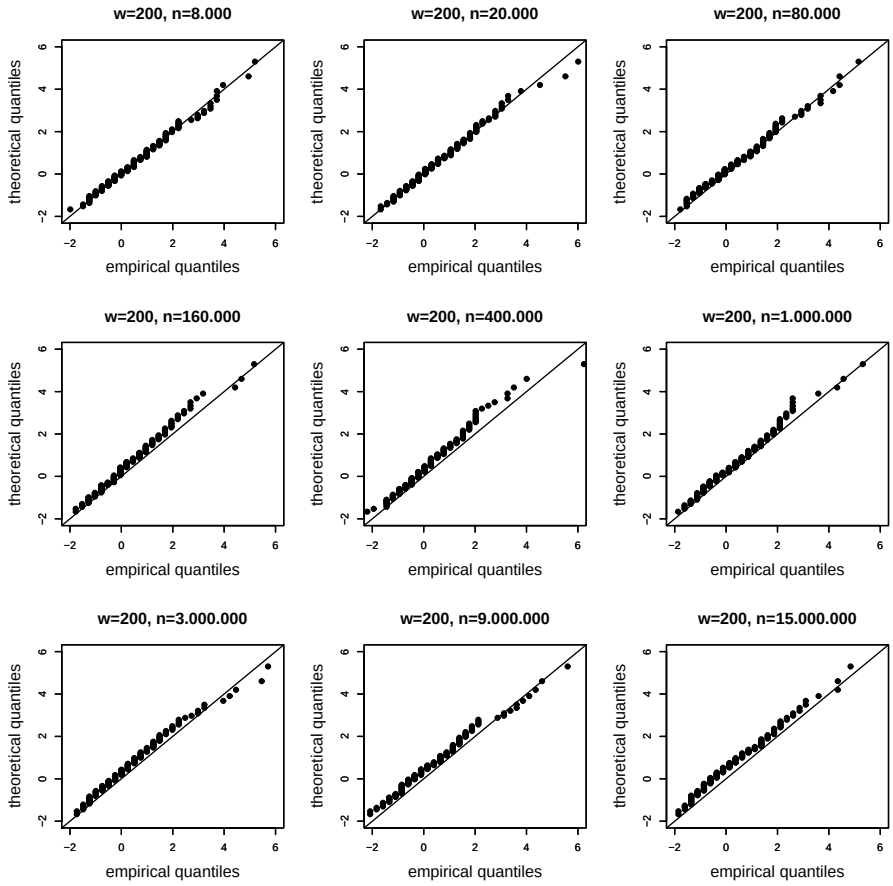

Figure 7.6: The figure shows under the Markov model the QQ-plots of the normalised empirical scores versus the theoretical Gumbel distribution for $w = 200$ and n ranging from 8000 to 15 millions.

Figure 7.5 and 7.6 show the corresponding results for the Markov chain simulation study. The conclusion is the same. The theory seems to fit the simulated data well in the wide range of n's from 400 to 15 millions keeping $w = 200$ fixed. The same tendency of the theory to slightly overestimate the maximal score is still observed for large n and for the same reasons.

7.5 Gaps and optimisation

It is natural to try to improve our search technique even further by introducing a finite internal gap penalty function. In that case we have no theory to support us. It is

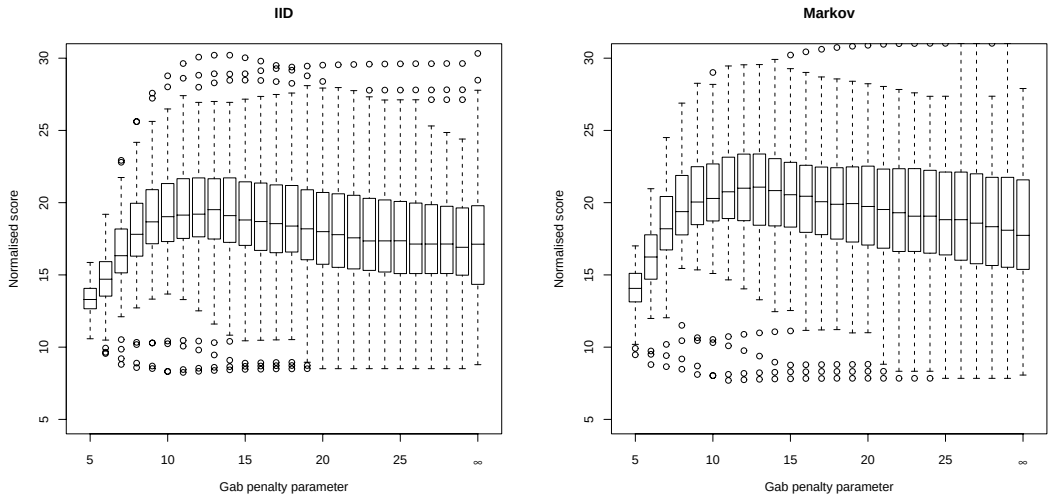

Figure 7.7: As we change the gap penalty parameter α, the (estimated) normalisation constants changes. Computing for each choice of the gap penalty parameter the normalised optimal score for each of the 59 miRNAs in the dataset results in a set of comparable distributions. Here we compare the boxplots using the *iid* model and *iid* scoring as well as the Markov model and Markov scoring. At the right hand side on each of the graphs we have the boxplot for the gapless empirical distribution, which corresponds to $\alpha = \infty$.

commonly believed for local alignments that the asymptotic Poisson approximation, and hence the asymptotic extreme value approximation, is still valid for certain finite gap penalty functions (Altschul et al. 1997). At least if the gap penalties are sufficiently large, cf. the discussion in Section 4.2.2. We believe that this is true for local structures also. In that case, even allowing for a finite internal gap penalty function, there exist constants θ^* and K^* such that we can normalise the maximal score to have a distribution that is asymptotically independent of the score function and the null hypothesis. Clearly, we don't have any analytic representations of these constants as we don't have any theory to support their existence. Nevertheless, it has been proposed (Altschul et al. 2001) that they can be estimated from simulations.

It will take us to far away from the main track of this thesis to go into a detailed (empirical) investigation of whether such an extrapolation of the theory is valid. There are also many interesting aspects related to the actual estimation of the parameters θ^* and K^* that we will not discuss. We refer to Altschul et al. (2001) and the reference therein for more details on these issues. Instead, we will try to give a hint about how much better we can actually do by allowing gaps. It is clear that the lower the gap cost is the higher the score achieved for the miRNAs will be. However, lowering

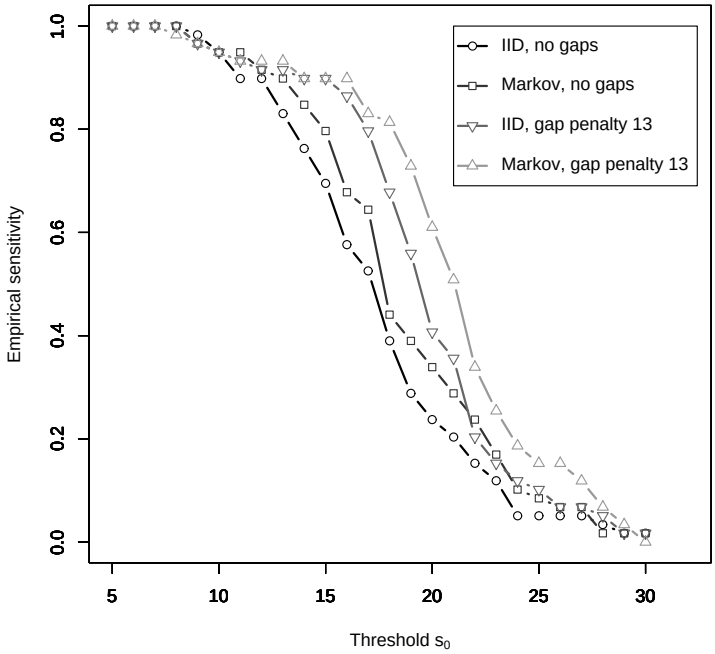

Figure 7.8: Choosing the gap penalty to be 13 gives for both *iid* and Markov scoring the maximal specificity with a 50% empirical sensitivity. Here we compare the empirical sensitivity as a function of the normalised threshold for gapped and ungapped scoring using either the *iid* or the Markov scoring.

the gap cost will also increase the maximal score in random sequences modelled by the null hypothesis. It seems plausible that the best performance is achieved for some 'intermediate' gap penalty function. What we will do is to introduce a linear internal gap penalty function $g_I(n) = \alpha n$ and investigate for different choices of α the effect of gaps on the normalised empirical scores for the miRNAs. We normalise by estimating $\theta^*(\alpha)$ and $K^*(\alpha)$ for each α using the approach described by Altschul et al. (2001). We let α be integer valued and running from 5 to 29. In Figure 7.7 we show, for both the *iid* and Markov case, the distributions of the normalised scores represented as boxplots for each value of the gap penalty parameter. Due to the normalisation these distributions are directly comparable. We observe that as α increases from 5 to around 10-13 the majority of the scores increase and thereafter gradually decrease again. For $\alpha \to \infty$ we get a distribution comparable with that of not allowing for gaps, as it is also shown in Figure 7.7. Depending of what we want to optimise, different gap penalties may be optimal. The boxplots show that there is little to gain by changing the threshold if we want the sensitivity to be close to 100%. On the other hand, if we want to optimise the median, which corresponds to obtaining the best specificity with a 50% empirical sensitivity, the figure shows that in both cases $\alpha = 13$ is optimal. If we want to optimise the 25% quantile, which corresponds to obtaining the best specificity with a 75% empirical sensitivity, we

find that for the *iid* case $\alpha = 11$ is optimal and for the Markov case $\alpha = 12$ is optimal. The difference between $\alpha = 11$, 12 and 13 is, however, negligible.

Figure 7.8 shows the empirical sensitivity function with no gaps and with $\alpha = 13$ for both the *iid* case and the Markov case. With or without gaps we see a slight improvement in going from the *iid* to the Markov scoring[2]. There is a larger effect in allowing for gaps and the best curve corresponds to the Markov scoring with gaps. It should be remarked that we haven't taken into account the optimism in the empirical sensitivity function. By optimism we mean the overestimation of the sensitivity $\beta^*_\kappa(s_0)$ by $\beta^*_\varepsilon(s_0)$ due to the fact that the score function has been constructed from the same dataset of miRNAs as the empirical sensitivity is computed from. This optimism might be larger for the Markov scoring than for the *iid* and this point is definitely worth investigating in more details.

7.6 Discussion

The purpose of this chapter was to illustrate how the theoretical results obtained can be applied and to what extend the asymptotic theory holds for finite n and w. By the simulation study we have illustrated that the theoretical results can indeed be applied for realistic choices of n and w. Moreover, we have illustrated the usefulness of the asymptotic theory by the ability to normalise the maximal scores in a way that make their distribution independent of the score function as well as the null hypothesis. In this way we can directly compare the performance of different score functions, say. This could also be done without the theory by using e.g. simulations, but this is very time consuming for large n, and it will always give results depending on a concrete choice of n and w. The theory provides us with a picture that is independent of n and w.

To turn the procedure presented here into a really useful search procedure for finding miRNA might require some more work. Indeed, applying for instance the Markov score function and a linear gap penalty function with $\alpha = 13$ and with a threshold chosen to be 18 we obtain an empirical sensitivity of 81%, thus we would locate 4 out of 5 miRNAs from the dataset. The (extrapolated) theory predicts, however, that searching a single chromosome of *C. elegans* with roughly 15 millions letters would still result in on average 45 random stem-loop structures with a normalised score exceeding 18. Actually running a search on a *C. elegans* chromosome results in many more findings due to at least the following two reasons; (i) other stem-loop structures that are present for other reasons, and (ii) repetitive patterns that by

[2]It is not entirely fair to make this comparison directly as we change score function as well as null hypothesis. In Section 7.3 we saw, however, that changing the null hypothesis and keeping the *iid* score function didn't really have an effect.

accident can form stem-loops. The first findings might be interesting in themselves whereas the second findings should somehow be filtered out.

There are several possibilities of improving the search by including e.g. evolutionary conservation of miRNA between species, or one could take advantage of the way that miRNA works by being complementary to parts of the mRNA. Such issues will hopefully be investigated in future research projects.

7.7 Notes

The miRNAs of C. elegans have been investigated thoroughly during the last years, and a detailed study employing some computational search procedures similar to the one presented here has been reported in Lim et al. (2003). The novel idea presented here is to apply the theoretical results for normalisation, which provides a means for comparing different search techniques. This method (or at least some further extension of the method) should be valuable for searching for other types of noncoding RNA as well, which possess other types of distinguishing structural features. As mentioned, miRNAs work by interfering with the translation of messenger RNA. An excellent and easily accessible review of the research in RNA interference and miRNAs can be found in the Scientific American paper "Censors of the Genome" by Lau & Bartel (2003).

A

Appendix

A.1 Computations with Laplace transforms

In this thesis – in particular in Chapter 5 – computations with and assumptions about the spectral radii of 'Φ-matrices' play a very important role. Most of these computations are easier to understand and come out much cleaner in the setup of comparing *iid* variables instead of Markov chains. In the *iid* setup, the spectral radii computations boil down to computations with Laplace transforms, which may be of independent interest. In addition, these computations allow us to comment on the conditions imposed by Siegmund & Yakir (2003) for a Poisson approximation to hold. By the arguments presented here, their condition can easily be generalised a little.

We consider two stochastic variables X and Y taking values in a set E, and we assume that $f : E \times E \to \mathbb{R}$ is a given function. Let the distribution of X be π_1 and the distribution of Y be π_2 and let $\pi = \pi_1 \otimes \pi_2$. We will not assume that E is finite, but that the Laplace transform

$$\varphi(\theta) = \mathbb{E}(\exp(\theta f(X,Y))) = \int \exp(\theta f(x,y))\pi(\mathrm{d}x, \mathrm{d}y) \tag{A.1}$$

of the distribution of $f(X,Y)$ exists (is $< \infty$) for all $\theta > 0$, that $\mu = \mathbb{E}(f(X,Y)) < 0$ and, furthermore, that $f(X,Y)$ takes positive values with positive probability. In this case, $\varphi(\theta) \to \infty$ for $\theta \to \infty$, and since $\partial_\theta \varphi(0) = \mu < 0$ there is a unique solution $\theta^* > 0$ to $\varphi(\theta) = 1$ due to convexity of φ. We define the measure π^* by

$$\frac{\mathrm{d}\pi^*}{\mathrm{d}\pi}(x,y) = \exp(\theta^* f(x,y)).$$

We let π_1^* and π_2^* denote the marginals of π^*.

A.1.1 Mean value inequalities

Under π^*, the mean

$$\mu^* = \int f(x,y)\pi^*(\mathrm{d}x,\mathrm{d}y) = \int f(x,y)\exp(\theta^* f(x,y))\pi(\mathrm{d}x,\mathrm{d}y)$$

is positive, and we ask how this mean relates to the mean of f under $\pi_1^* \otimes \pi_2^*$ as well as under $\pi_1^* \otimes \pi_2$ or $\pi_1 \otimes \pi_2^*$.

Introducing the Laplace transform[1]

$$\varphi_\infty^*(\theta) = \int \exp(\theta(f(x,z)+f(w,y)-f(x,y)-f(w,z)))\pi^* \otimes \pi^*(\mathrm{d}x,\mathrm{d}y,\mathrm{d}w,\mathrm{d}z),$$

we see that $\varphi_\infty^*(0) = \varphi_\infty^*(\theta^*) = 1$, and with

$$\mu_\infty^* = \int f(x,y)\pi_1^* \otimes \pi_2^*(\mathrm{d}x,\mathrm{d}y)$$

we obtain $\partial\varphi_\infty^*(0) = 2\mu_\infty^* - 2\mu^*$. Hence if just

$$\mathbb{P}(f(X,Z)+f(W,Y) \neq f(X,Y)+f(W,Z)) > 0 \tag{A.2}$$

with $\mathcal{D}(X,Y,W,Z) = \pi \otimes \pi$, the Laplace transform φ_∞^* is strictly convex implying that $\mu_\infty^* < \mu^*$. Compare this result with Lemma 5.5.8.

We observe that if E is finite and $\pi_1(x) > 0$ and $\pi_2(x) > 0$ for all $x \in E$, (A.2) is equivalent to the existence of $x, y, z, w \in E$ such that

$$f(x,z)+f(w,y) \neq f(x,y)+f(w,z). \tag{A.3}$$

If this is *not* the case we can fix some $w_0, z_0 \in E$ such that for all $x, y \in E$

$$f(x,y) = f(x,z_0) - f(w_0,z_0) + f(w_0,y) = f_1(x) + f_2(y)$$

with e.g. $f_1(x) = f(x,z_0) - f(w_0,z_0)$ and $f_2(y) = f(w_0,y)$. On the other hand, if $f(x,y) = f_1(x) + f_2(y)$ for some f_1 and f_2 then (A.2) clearly doesn't hold.

Likewise, we can consider the Laplace transform

$$\tilde{\varphi}^*(\theta) = \int \exp(\theta(f(x,z)-f(x,y)))\pi^* \otimes \pi_2(\mathrm{d}x,\mathrm{d}y,\mathrm{d}z),$$

for which $\tilde{\varphi}^*(0) = \tilde{\varphi}^*(\theta^*) = 1$, and with

$$\tilde{\mu}^* = \int f(x,y)\pi_1^* \otimes \pi_2(\mathrm{d}x,\mathrm{d}y)$$

[1] with notation φ_∞^* chosen so that it is comparable with the notation in Section 5.5.4

we have that $\partial\tilde{\varphi}^*(0) = \tilde{\mu}^* - \mu^*$. So if

$$\mathbb{P}(f(X,Z) \neq f(X,Y)) > 0$$

with $\mathcal{D}(X,Y,Z) = \pi \otimes \pi_2$, the Laplace transform $\tilde{\varphi}^*$ is strictly convex, hence $\tilde{\mu}^* < \mu^*$. Compare with Lemma 5.5.7. In the *iid* setup there is no result corresponding to Lemma 5.5.9.

A.1.2 Two-dimensional Laplace transforms

The idea to consider the Laplace transforms of the previous section actually comes from a two-dimensional Laplace transform point of view. A two-dimensional point of view that is closely related to Condition 5.9. With

$$\varphi_1(\theta,\eta) = \int \exp(\theta f(x,y) + \eta f(x,z))\pi \otimes \pi_2(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z),$$

we observe that $\varphi_1(\theta,0) = \varphi_1(0,\theta) = \varphi(\theta)$. Thus in particular,

$$\varphi_1(0,0) = \varphi_1(\theta^*,0) = \varphi_1(0,\theta^*) = 1. \tag{A.4}$$

The φ_1-spectral radius appearing in Condition (5.9) corresponds to $\varphi_1(\theta,\theta)$ in this setup. Differentiation yields that

$$\nabla\varphi_1(0,\theta^*) = (\tilde{\mu}^*, \mu^*),$$

and since $\nabla\varphi_1(0,\theta^*)$ is orthogonal to the tangent at $(0,\theta^*)$ of the convex contour curve given by $\varphi_1(\theta,\eta) = 1$, we get from (A.4) that $\tilde{\mu}^* \leq \mu^*$. To get a sharp inequality, one needs to know whether φ_1 is strictly convex. This two-dimensional, geometric argument has been boiled down to a one-dimensional Laplace transform argument using $\tilde{\varphi}^*$ above. In fact, $\tilde{\varphi}^*$ is equal to φ_1 evaluated on the line given by $\eta = \theta^* - \theta$.

A corresponding two-dimensional point of view for discovering φ_∞^* is of course possible. We skip the details and instead work out more consequences of the two-dimensional Laplace transform φ_1.

With e a unit vector in $[0,\infty)^2$ (in the 1-norm) we let $\gamma_e^* > 0$ be the solution of $\varphi_1(\gamma e) = 1$ (assuming that there is such a solution). We observe that for $e = (0,1)$ and $e = (1,0)$ we have $\gamma_e^* = \theta^*$. Moreover, due to convexity of the contour curve given by $\varphi_1(\theta,\eta) = 1$, we have that $\gamma_e^* \geq \theta^*$ for all e. In particular, for $e = (1/2, 1/2)$ we see that $\gamma_e^* \geq \theta^*$ corresponding to $\varphi_1(\theta^*/2, \theta^*/2) \leq 1$. Compare with Condition (5.9).

Let $(X_n)_{n\geq 1}$, $(Y_n)_{n\geq 1}$ and $(Z_n)_{n\geq 1}$ be independent sequences of *iid* variables such that the X's have distribution π_1 and the Y's and Z's have distribution π_2, and define

$$\begin{aligned} S_1 &= \sum_{k=1}^{n} f(X_n, Y_n) \quad \text{and} \\ S_2 &= \sum_{k=1}^{n} f(X_n, Z_n). \end{aligned}$$

For any $s \geq 0$ and any unit vector e it follows by an exponential change of measure that

$$\mathbb{P}(S_1 \geq s, S_2 \geq s) \leq \exp(-s\gamma_e^*).$$

Due to convexity (again) and symmetry, the maximal γ_e^* is found for $e = (1/2, 1/2)$. Thus using this exponential change of measure technique we obtain the best possible two-dimensional, exponential inequality by considering φ_1 on the diagonal (θ, θ). This is the rational behind the assumption made in Condition (5.9) – we can not do better by making other, non-symmetric exponential changes of measure. We also see in the light of the exponential inequality above, that (5.9) really is an assumption about the simultaneous excess of a level s when comparing two sequences, one by one, with a third using f.

Finally, one should relate the derivations presented here with Theorem 3 in Siegmund & Yakir (2003) and their condition for a Poisson approximation to hold. Without referring to the two-dimensional Laplace transform setup presented here, they choose to consider a not necessarily optimal exponential change of measure. They assume that $\varphi(\theta^*, \eta) = 1$ has a solution $\eta^* > 0$ and choose to consider the direction $e \propto (\theta^*, \eta^*)$. Furthermore, they essentially assume (in the case $n = m$ in their notation) that $\eta^* > \theta^*/2$, which implies by a convexity argument that $\varphi_1(3/4\theta^*, 3/4\theta^*) < 1$. Their results are actually valid under the more general assumption that $\varphi_1(3/4\theta^*, 3/4\theta^*) < 1$ holds (if $n = m$).

Bibliography

Alsmeyer, G. (1994), 'On the Markov renewal theorem', *Stochastic Process. Appl.* **50**(1), 37–56.

Altschul, S. F., Bundschuh, R., Olsen, R. & Hwa, T. (2001), 'The estimation of statistical parameters for local alignment score distributions', *Nucleic Acids Research* **29**(2), 351–361.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997), 'Gapped blast and psi-blast: a new generation of protein database search programs', *Nucleic Acids Research* **25**(17), 3389–3402.

Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990), 'Basic local alignment search tool', *J. Mol. Biol.* **215**, 403–410.

Arratia, R., Goldstein, L. & Gordon, L. (1989), 'Two moments suffice for Poisson approximations: the Chen-Stein method', *Ann. Probab.* **17**(1), 9–25.

Arratia, R. & Waterman, M. S. (1994), 'A phase transition for the score in matching random sequences allowing deletions', *Ann. Appl. Probab.* **4**(1), 200–225.

Asmussen, S. (1998), 'Subexponential asymptotics for stochastic processes: extremal behavior, stationary distributions and first passage probabilities', *Ann. Appl. Probab.* **8**(2), 354–374.

Asmussen, S. (2003), *Applied probability and queues*, Vol. 51 of *Applications of Mathematics*, second edn, Springer-Verlag, New York. Stochastic Modelling and Applied Probability.

Asmussen, S., Kalashnikov, V., Konstantinides, D., Klüppelberg, C. & Tsitsiashvili, G. (2002), 'A local limit theorem for random walk maxima with heavy tails', *Statist. Probab. Lett.* **56**(4), 399–404.

Asmussen, S. & Perry, D. (1992), 'On cycle maxima, first passage problems and extreme value theory for queues', *Comm. Statist. Stochastic Models* **8**(3), 421–458.

C. Lee, R. & Ambros, V. (2001), 'An extensive class of small RNAs in Caenorhabditis elegans', *Science* **294**, 862–864.

Chen, L. H. Y. (1975), 'Poisson approximation for dependent trials', *Ann. Probability* **3**(3), 534–545.

Çinlar, E. (1975), *Introduction to stochastic processes*, Prentice-Hall Inc., Englewood Cliffs, N.J.

Dembo, A., Karlin, S. & Zeitouni, O. (1994*a*), 'Critical phenomena for sequence matching with scoring', *Ann. Probab.* **22**(4), 1993–2021.

Dembo, A., Karlin, S. & Zeitouni, O. (1994*b*), 'Limit distribution of maximal non-aligned two-sequence segmental score', *Ann. Probab.* **22**(4), 2022–2039.

Doukhan, P. (1994), *Mixing*, Springer-Verlag, New York. Properties and examples.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997), *Modelling extremal events*, Vol. 33 of *Applications of Mathematics*, Springer-Verlag, Berlin. For insurance and finance.

Ewens, W. J. & Grant, G. R. (2001), *Statistical methods in bioinformatics: an introduction*, Statistics for Biology and Health, Springer-Verlag, New York.

Foss, S. & Zachary, S. (2003), 'The maximum on a random time interval of a random walk with long-tailed increments and negative drift', *Ann. Appl. Probab.* **13**(1), 37–53.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. (2003), 'Rfam: an RNA family database', *Nucleic Acids Research* **31**(1), 439–441.

Hansen, N. & Jensen, A. T. (2003), The extremal behaviour over regenerative cycles for markov additive processes with heavy tails, Preprint 1, Department of Applied Mathematics and Statistics, University of Copenhagen.

Heath, D., Resnick, S. & Samorodnitsky, G. (1997), 'Patterns of buffer overflow in a class of queues with long memory in the input stream', *Ann. Appl. Probab.* **7**(4), 1021–1057.

Iglehart, D. L. (1972), 'Extreme values in the $GI/G/1$ queue', *Ann. Math. Statist.* **43**, 627–635.

Jelenković, P. R. & Lazar, A. A. (1998), 'Subexponential asymptotics of a Markov-modulated random walk with queueing applications', *J. Appl. Probab.* **35**(2), 325–347.

Karlin, S. & Dembo, A. (1992), 'Limit distributions of maximal segmental score among Markov-dependent partial sums', *Adv. in Appl. Probab.* **24**(1), 113–140.

Kingman, J. F. C. (1961), 'A convexity property of positive matrices', *Quart. J. Math. Oxford Ser. (2)* **12**, 283–284.

Klüppelberg, C. (1988), 'Subexponential distributions and integrated tails', *J. Appl. Probab.* **25**(1), 132–141.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. (2001), 'Identification of novel genes coding for small expressed RNAs', *Science* **294**(October 26), 853–858.

Lau, N. C. & Bartel, D. P. (2003), 'Censors of the genome', *Scientific American* **289**(2), 26–33.

Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. (2001), 'An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans', *Science* **294**, 858–862.

Ledoux, M. & Talagrand, M. (1991), *Probability in Banach spaces*, Vol. 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*, Springer-Verlag, Berlin. Isoperimetry and processes.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B. & Bartel, D. P. (2003), 'The microRNAs of Caenorhabditis elegans', *Genes Dev.* **17**, 991–1008.

Neveu, J. (1965), *Mathematical foundations of the calculus of probability*, Translated by Amiel Feinstein, Holden-Day Inc., San Francisco, Calif.

O'Cinneide, C. (2000), 'Markov additive processes and Perron-Frobenius eigenvalue inequalities', *Ann. Probab.* **28**(3), 1230–1258.

Petrov, V. V. (1995), *Limit theorems of probability theory*, Vol. 4 of *Oxford Studies in Probability*, The Clarendon Press Oxford University Press, New York. Sequences of independent random variables, Oxford Science Publications.

Siegmund, D. & Yakir, B. (2000), 'Approximate p-values for local sequence alignments', *Ann. Statist.* **28**(3), 657–680.

Siegmund, D. & Yakir, B. (2003), 'Correction: "Approximate p-values for local sequence alignments" [Ann. Statist. **28** (2000), no. 3, 657–680; MR 2002a:62140]', *Ann. Statist.* **31**(3), 1027–1031.

Takahata, H. (1981), 'L_∞-bound for asymptotic normality of weakly dependent summands using Stein's result', *Ann. Probab.* **9**(4), 676–683.

Wald, A. (1947), *Sequential Analysis*, John Wiley & Sons Inc., New York.

Waterman, M. S., ed. (1995), *Introduction to Computatinal Biology*, Chapman & Hall/CRC, Boca Raton, Florida.

Yakir, B. & Grossmann, S. (2001), 'Large deviations for optimal local seqeunce alignment', *http://pluto.mscc.huji.ac.il/∼msby/index.html* .

Zuker, M. (2003), 'Mfold web server for nucleic acid folding and hybridization prediction.', *Nucleic Acids Research* **31**(13), 3406–3415.