Statistics of optimal sequence alignments

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Mathematik
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Steffen Grossmann
aus Offenbach am Main

Frankfurt 2003
(D F 1)

vom Fachbereich Mathematik der
Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan:	Prof. Dr. J. Baumeister
Gutachter:	Prof. Dr. A. Wakolbinger Prof. Dr. N. Gantert
Datum der Disputation:	7. Mai 2003

Preface

This thesis has been motivated by the problem of assessing the statistical significance of the outcomes from sequence alignment algorithms. Finding similarities between sequences with entries from a finite alphabet has become an important task in recent times — especially in biology, where a fast growing number of sequences from the 4-letter DNA-alphabet or the 20-letter amino acid-alphabet waits to be analyzed and understood.

In this thesis we are concerned with a class of similarity criteria, that is widely used in biology. It expresses the similarity of two sequences by giving an *optimized local alignment*. A local alignment basically gives the information *which* parts of the sequences are similar and *what* the similarity looks like in detail. This local alignment is optimized in the following sense. First a score is assigned to each possible local alignment of the two sequences. Here, the scoring scheme should be designed in such a way that scores increase with the degree of similarity in the alignment. Then the maximizing alignment is determined. This alignment, together with its score, measures the degree of similarity between the two sequences.

We will see shortly that the scoring schemes are chosen in such a way that determining the optimal alignment can be done in reasonable time. However there remains the problem to understand the similarity scale that is given by this procedure: *How high does a score have to be in order to indicate that the two sequences share some extraordinary similarity?* A natural way to answer this question would be to determine which scores one would observe when comparing two *typical unrelated* sequences. If a score observed from real data is very high compared to those typical scores, then this would indicate a strong similarity.

Put more mathematically, the term *typical unrelated* has to be read as *random independent*. Thus, the statistical way to answer the above question, is to take two sequences of independently chosen random letters and determine the distribution of the optimal local alignment score.

Although this sounds simple, this is by no means the case. The optimal local alignment score is a complicated quantity, since in its calculation a maximization over the set of *all possible* local alignments of the two sequences is involved. This makes it difficult to calculate its distribution. In fact, from the point of view of probability theory, the resulting model is

closely related to so-called *percolation models* — a class of models that has been actively researched during the past two decades, but for which many challenging problems still wait for their solution.

Outline

In Chapter 1 we start with explaining in more detail how optimal alignments are defined. Their use in biology will be briefly described. To understand why alignments are important for biologists it is also necessary to explain some of the more theoretical biological concepts behind them. Throughout the thesis we will repeatedly see, why the model is interesting from the mathematical point of view. Here we start by explaining the relationship between alignments and percolation theory. This helps understanding the complexity of the problem. Also the statistics of optimal local alignments lead to questions which in that form do not appear in percolation theory but are crucial for this thesis.

In order to better understand the main results of this thesis we give an extended overview in Chapter 2. The exposition here is heuristical and does not claim to fulfil the requirements of mathematical precision. Its objective is to give a readable but yet profound insight into the mathematical ideas used. Nevertheless we always make clear where in the later chapters the mathematically precise statements can be found.

In Chapter 3 we leave the world of alignments for a moment and give our main general result, namely, we characterize the large deviations of the global maximum of an independent superadditive process with negative drift. We formulate it in a language that is inspired by large deviations theory of random walks, a setting which provides the right general framework.

Results specific to the model of optimal alignments are given in Chapter 4. These comprise a mathematical precise definition of the model, concentration inequalities for the optimal global score, large deviations results for optimal global and local scores, a phase transition result for the mean behaviour of the optimal local score and several results about convergence rates. All the results here and in Chapter 3 are presented in a mathematically precise way, complementing the heuristical exhibition in Chapter 2.

Finally, we end in Chapter 5, discussing the results and practical issues related to our results. Also we give an outlook on possible further work.

Acknowledgements

The first person I want to thank is my supervisor Prof. Dr. Anton Wakolbinger from Johann Wolfgang Goethe-University, Frankfurt. During the whole process of doing my PhD, he was there when I needed his advice or

his experience. Also, he is the one who motivated me to present the results at several workshops and conferences.

This thesis would not be what it is without Dr. Benjamin Yakir from Hebrew University, Jerusalem, Israel. I want to thank him for teaching me how to ask the right questions. His question in how far the large deviations rates of the optimal local alignment scores differ in the gapless and gapped case was the spark that enabled me to get the results presented in this thesis. To the results presented in Sections 4.9 and 4.10 he even contributed actively: These we derived together at several mutual visits in Jerusalem and Frankfurt.

During the past years I have enjoyed much the working atmosphere at the probability group of Frankfurt University's Math Department. Among my colleagues I especially want to thank Dr. Brooks Ferebee, Dr. Dirk Metzler and Dr. habil. Jochen Geiger, who has meanwhile migrated to Kaiserslautern.

Finally, I want to thank my family for the good times we have had during the past years and my parents for their never-ending and unconditional support.

Contents

List of Figures

Statistik optimaler Sequenzalignments

Zusammenfassung

Diese Dissertation beschäftigt sich mit statistischen Fragen optimaler Sequenzalignments. Optimale Sequenzalignments sind ein wichtiges Werkzeug bei der automatisierten Suche nach Ähnlichkeiten zwischen DNA- oder Aminosäuresequenzen. Im Labor neu bestimmte Sequenzen werden heutzutage routinemäßig mit großen Datenbanken bekannter Sequenzen verglichen. Die dabei verwendeten Programme bestimmen ein optimales lokales Alignment zwischen der neuen Sequenz und der gesamten Datenbank zusammen mit dem Score des Alignments. Der ermittelte Score ist ein Maß für die Güte des Alignments — je höher der Score, desto besser das Alignment.

Bei einer solchen Datenbanksuche stellt sich direkt die Frage nach der Normierung dieser durch den Alignmentscore definierten Ähnlichkeitsskala. Selbst wenn alle beteiligten Sequenzen rein zufällig erzeugt wären, und demnach keine systematischen Ähnlichkeiten aufwiesen, würden hohe Scorewerte oftmals auch per Zufall erzielt werden. Wie hoch muss also ein Score sein, damit man annehmen kann, dass er nicht auch schon durch Zufall hätte erzielt werden können? Oder, präziser ausgedrückt: Wie wahrscheinlich ist es unter dem Nullmdell unabhängiger, zufälliger Sequenzen, optimale lokale Scores zu erhalten, die eine vorgegebene hohe Schranke übersteigen? Diese Wahrscheinlichkeit wird auch als der *p-Wert* der Schranke bezeichnet.

Dieser Frage geht diese Dissertation nach. Dabei untersuchen wir speziell den Fall optimaler lokaler Alignments, in denen auch Lücken erlaubt sind.

Optimale Alignments

Ein Beispiel eines Alignments zweier kurzer DNA-Sequenzen findet sich im oberen Teil von Abb. 1.1 auf Seite 2. Die im Alignment übereinander stehenden Positionen der Sequenzen werden als zueinander gehörig angesehen, während mit Hilfe von Lücken die Positionen ausgerichtet werden, für die es in der jeweils anderen Sequenz keine Entsprechung gibt.

Das Ziel ist es, ein möglichst gutes Alignment zweier gegebener Sequenzen zu finden. Dazu berechnet man für jedes mögliche Alignment der beiden Sequenzen wie folgt einen speziellen, additiven Score. Man gibt sich eine Scoringmatrix K vor, die für jede mögliche Kombination zweier Buchstaben einen Eintrag besitzt. Aus dieser Matrix liest man für jedes alignierte Buchstabenpaar den entsprechenden Wert ab und summiert diese Werte auf. Üblicherweise besitzt die Scoringmatrix für Paare gleicher Buchstaben positive und für Paare ungleicher Buchstaben negative Einträge. Die Scoringmatrix kann dabei so gewählt werden, dass sie gewisse Buchstabenkombinationen in den Alignments mehr bevorzugt als andere. Der bis jetzt berechnete Scorewert wird nun noch verkleinert, um die Anzahl der Lücken in dem Alignments nicht zu groß werden zu lassen. Hierzu gibt man sich zwei Parameter $\delta \geq 0$ und $\Delta \geq 0$ vor und subtrahiert für jede Lücke im Alignment einmal den Wert δ und für jede Gruppe von Lücken einmal den Wert Δ. Die Parameter Δ bzw. δ werden deshalb auch als *Gap-opening-* bzw. *Gap-extension-penalty* bezeichnet.

Ein optimales Alignment ist nun ein Alignment, welches die so definierte Scoringfunktion maximiert. Hierbei unterscheidet man zwischen optimalen *globalen* und optimalen *lokalen* Alignments. Ein globales Alignment muss alle Positionen der beiden gegebenen Sequenzen verwenden (wobei natürlich nicht alle Positionen in alignierten Paarungen verwendet werden müssen), während in einem lokalen Alignment nur die Positionen aus jeweils einem beliebigen Teilintervall der beiden Sequenzen beteiligt sein müssen. Die oben beschriebenen Datenbanksuchverfahren basieren immer auf lokalen Alignments, da nur diese in der Lage sind Ähnlichkeiten zwischen Teilen der beiden verglichenen Sequenzen zu finden.

Auch wenn es für zwei gegebene Sequenzen eine sehr große Anzahl möglicher Alignments gibt, so gibt es doch Algorithmen, die in der Lage sind, ein optimales lokales oder globales Alignment in einer Laufzeit zu bestimmen, die proportional zum Produkt der Sequenzlängen ist. Diese Algorithmen basieren auf Prinzipien der Dynamischen Programmierung und sind unter den Namen *Needleman-Wunsch-* (optimale globale Alignments) und *Smith-Waterman-Algorithmus* (optimale lokale Alignments) bekannt.

Lokales und globales Regime

Im folgenden bezeichnen wir mit S_n bzw. M_n den optimalen globalen bzw. lokalen Score zweier zufälliger Sequenzen der Länge n, wobei alle Buchstaben der Sequenzen unabhängig gemäß einer gegebenen Verteilung μ auf dem zugrunde liegenden Alphabet $\mathcal{A}$ gezogen wurden. Das Verhalten von M_n ist sehr stark von den gewählten Scoringparametern abhängig. Aus Superadditivitätsargumenten folgt zunächst

$$\lim_{n\to\infty} \frac{\mathbb{E}[S_n]}{n} = \sup_{n>0} \frac{\mathbb{E}[S_n]}{n} =: \gamma \in \mathbb{R} \tag{1}$$

und es stellt sich heraus, dass das Vorzeichen von γ entscheidend für das Verhalten von $\mathbb{E}[M_n]$ ist. Im Fall $\gamma > 0$ verhält sich $\mathbb{E}[M_n]$ wie $\mathbb{E}[S_n]$ — der interessante Fall ist $\gamma < 0$, denn dies bedeutet, dass optimale globale Alignments langer Sequenzen im Mittel sehr schlechte Scores erzielen. In diesem Fall besteht bei lokalen Alignments die Möglichkeit, einen hohen positiven Score zu erzielen, wenn zumindest Teile der beiden Sequenzen einander ähnlich sind. In der Tat wächst $\mathbb{E}[M_n]$ in diesem Fall logarithmisch in n, genauer gesagt gibt es eine Konstante $b > 0$ mit

$$b = \lim_{n\to\infty} \frac{\mathbb{E}[M_n]}{2\log n}. \tag{2}$$

Diese zwei Phasen im Verhalten von M_n wurden zuerst in [AW94] beschrieben. Für die beiden Phasen haben sich die Bezeichnungen *globales Regime* ($\gamma > 0$) bzw. *lokales Regime* ($\gamma < 0$) eingebürgert.

Die bei der Suche nach lokalen Ähnlichkeiten in Sequenzdatenbanken verwendeten Scoringparameter müssen sinnvollerweise aus dem lokalen Regime gewählt sein. Hierbei werden für die Buchstabenverteilung μ des Nullmodells üblicherweise die in der Datenbank auftretenden relativen Häufigkeiten der Buchstaben genommen.

Die weiter unten diskutierten Ergebnisse dieser Dissertation ermöglichen eine viel besser strukturierte Sicht auf dieses Phasenübergangsresultat als in [AW94]. Wir diskutieren es deshalb ausführlich in dieser Dissertation, insbesondere geben wir einen vereinfachten Beweis.

Konzentrationsungleichungen

Nachdem wir nun wissen, dass $\mathbb{E}[S_n]$ aus Superadditivitätsgründen asymptotisch linear wächst, stellt sich die Frage, wie stark die Verteilung von S_n um ihren Erwartungswert konzentriert ist. Abschätzungen für solche Streuungen werden gemeinhin als *Konzentrationsungleichungen* bezeichnet.

Für S_n wurde die entsprechende Konzentrationsungleichung bereits in [AW94] bewiesen. Sie besagt, dass eine Konstante c existiert für die gilt

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\exp\left(-\frac{t^2}{8c^2 n}\right). \tag{3}$$

Dies wurde in [AW94] unter Benutzung der Azuma-Hoeffding-Ungleichung bewiesen. Wir zeigen, dass sich eine vergleichbare, wenn auch geringfügig schwächere Konzentrationsungleichung auch mit Hilfe einer isoperimetrischen Ungleichung beweisen lässt — einer grundlegend anderen Methoden, die in den letzten Jahren von M. Talagrand (vgl. [Tal96]) entwickelt wurden.

Datenbankvergleich und Vergleich zweier Sequenzen

Im folgenden beschränken wir uns auf Scoringschemata und Buchstabenverteilungen aus dem lokalen Regime.

Rekapitulieren wir noch einmal die zugrunde liegende Fragestellung. Bezeichnen wir dazu den bei der Bestimmung des optimalen lokalen Alignments zwischen einer neuen Sequenz und den Sequenzen einer großen Datenbank erzielten Score mit $\bar{M}$ und nehmen wir an, dass er einen Wert t erreicht hat. Aus der Sicht der Statistik müssen wir $\mathbb{P}(\bar{M} \geq t)$ unter dem Nullmodell zufälliger, unabhängiger Sequenzen berechnen. Falls diese Wahrscheinlichkeit sehr klein ist, so würde man die Nullhypothese, dass keine signifikante, lokale Ähnlichkeit zwischen der neuen Sequenz und irgendeinem Teil der Datenbank vorliegt, verwerfen und sich das gefundene lokale Alignment genauer ansehen.

$\bar{M}$ selbst ist das Maximum von allen Scores, die bei den einzelnen Vergleichen der neuen Sequenz mit den Sequenzen der Datenbank erzielt wurden. Unter dem Nullmodell ist $\bar{M}$ also das Maximum einer gewissen Anzahl N unabhängiger optimaler lokaler Scores aus dem Vergleich jeweils zweier Sequenzen. Nehmen wir der Einfachheit halber einmal an, dass alle beteiligten Sequenzen von der gleichen Länge n sind, so sind die N unabhängigen Scores, über die maximiert wird, alle wie M_n verteilt.

Aus der Extremwerttheorie ist nun bekannt, dass das asymptotische Verhalten eines Maximums von unabhängig, identisch verteilten Zufallsvariablen vor allem durch die Asymptotik des oberen Schwanzes der Verteilung der einzelnen Zufallsvariablen bestimmt wird. Mit anderen Worten, um $\mathbb{P}(\bar{M} \geq t)$ zu berechnen, muss man vor allem wissen wie sich $\mathbb{P}(M_n \geq t)$ für große n und t verhält.

Alignments ohne Lücken

Als Ausgangspunkt für die Beschreibung unserer Ergebnisse zur Verteilung von M_n im allgemeinen Fall *mit* Lücken, beschreiben wir zunächst den viel einfacheren Fall optimaler lokaler Alignments *ohne* Lücken. Dieser lässt sich als ein Spezialfall allgemeiner Alignments mit Lücken auffassen, indem man z. B. $\Delta = \infty$ setzt. Er wurde bereits 1994 von Dembo, Karlin und Zeitouni (vgl. [DKZ94a] und [DKZ94b]) vollständig behandelt.

Dort wurde u. a. gezeigt, dass zwei Konstanten θ^* und K^* existieren, sodass gilt

$$\lim_{n\to\infty} \mathbb{P}\left(M_n - \frac{2\log n}{\theta^*} \leq t\right) = \exp(-K^* \exp(-\theta^* t)). \tag{4}$$

Die Verteilungsfunktion auf der rechten Seite ist aus der klassischen Extremwerttheorie als *Extremwertverteilung des Typs I* (auch *Gumbel-Verteilung*) bekannt. Sie taucht als Grenzverteilung des Maximums mehrerer unabhängiger, identisch verteilter Zufallsvariablen auf, falls diese eine Verteilung mit einem exponentiell abfallenden (oberen) Schwanz besitzen.

Daran orientiert sich auch der Beweis von Dembo, Karlin und Zeitouni. Heuristisch formuliert zeigen sie, dass im Fall ohne Lücken M_n aufgefasst

werden kann als das Maximum von n^2 asymptotisch unabhängigen Zufallsvariablen mit identischer Verteilung, für welche der obere Schwanz asymptotisch wie $K^* \exp(-\theta^* t)$ abfällt.

Für uns wichtig ist hierbei vor allem die Charakterisierung der Konstanten θ^*. Bezeichnen wir mit $\Lambda(\lambda)$ die logarithmierte momenterzeugende Funktion des Scores eines unter dem Nullmodell zufällig gewählten Buchstabenpaares, so ist θ^* gegeben durch die eindeutig bestimmte, positive Lösung von

$$\log \mathbb{E}[\exp(\lambda K(X,Y))] =: \Lambda(\lambda) = 0. \tag{5}$$

Für den Schwanz der Verteilung von M_n lässt sich (4) lesen als

$$\mathbb{P}(M_n \geq t) \sim n^2 K^* \exp(-\theta^* t). \tag{6}$$

Dies ist die eigentliche Aussage, die benötigt wird, um die oben formulierte, statistische Fragestellung in der Praxis zu beantworten. Es wäre wünschenswert, eine ebenso genaue Kenntnis der Asymptotik des Schwanzes der Verteilung von M_n für den Fall optimaler lokaler Alignments mit Lücken zu besitzen. Die Resultate der vorliegenden Arbeit sind ein Schritt in Richtung dieses Ziels.

Große Abweichungen

Auf der Ebene der großen Abweichungen bedeutet (6), dass

$$\lim_{n,t\to\infty} -\frac{1}{t} \log \mathbb{P}(M_n \geq t) = \theta^*$$

gilt, solange $t = o(n)$ und $\log n = o(t)$ erfüllt ist. Eines unserer Hauptergebnisse ist, dass dieses Ergebnis auch für den Fall optimaler lokaler Alignments mit Lücken gilt. Wir zeigen außerdem, dass sich die oben beschriebene Charakterisierung von θ^* in natürlicher Weise auf den Fall mit Lücken fortsetzen lässt.

Hierzu definieren wir zunächst für jedes n die logarithmierte momenterzeugende Funktion von S_n als

$$\Lambda_n(\lambda) := \log \mathbb{E}[\exp(\lambda S_n)].$$

Aus Superadditivitätsargumenten folgt für $\lambda \geq 0$ die Existenz der Grenzfunktion

$$\Lambda(\lambda) := \sup_{n>0} \frac{1}{n} \Lambda_n(\lambda) = \lim_{n\to\infty} \frac{1}{n} \Lambda_n(\lambda). \tag{7}$$

Bei Alignments ohne Lücken stimmt diese Definition von Λ mit der Definition in (5) überein, denn es gilt $(1/n)\Lambda_n = \Lambda$ für alle n. Im allgemeinen Fall mit Lücken jedoch bilden die Funktionen $(1/n)\Lambda_n$ eine echt aufsteigende Folge und sind alle von der Grenzfunktion Λ verschieden.

Unser zentrales Theorem ist nun

Theorem D.1. *Es sei θ^* die eindeutig bestimmte, positive Lösung von*

$$\Lambda(\lambda) = 0,$$

dann gilt

$$\lim_{n,t\to\infty} -\frac{1}{t}\log \mathbb{P}(M_n \geq t) = \theta^*,$$

solange $t = o(n)$ und $\log n = o(t)$ erfüllt sind.

Desweiteren zeigen wir, dass die in (2) definierte Wachstumskonstante b des logarithmischen Wachstums von $\mathbb{E}[M_n]$ mit dem eben definierten θ^* durch die Gleichung

$$b = \frac{1}{\theta^*}$$

verbunden ist.

Der Beweis von Theorem D.1 beruht zum einen auf Superadditivitätsargumenten, sowie auf Konvexitätsargumenten. Dabei kommt ein bereits 1974 von Hammersley in [Ham74] bewiesenes Ergebnis über Große Abweichungen unabhängiger, superadditiver Prozesse zum Einsatz. Wir diskutieren dieses Ergebnis ausführlich, besonders den Zusammenhang zu dem in der jüngeren Theorie der Großen Abweichungen wichtigen Gärtner-Ellis-Theorem. Desweiteren benutzen wir die oben beschriebene Konzentrationsungleichung (3), die gewisse Eigenschaften der Grenzfunktion Λ sichert.

Konvergenzraten

Im Zusammenhang mit Superadditivitätsargumenten für optimale globale Alignments sind uns bis jetzt zwei Grenzübergänge begegnet. Zum einen in der Definition (1) von γ, zum anderen bei der Definition (7) der Grenzfunktion Λ. Aufgrund einer speziellen Zerlegungseigenschaften optimaler globaler Alignments ist es möglich, in beiden Fällen genauere Aussagen über die Konvergenzgeschwindigkeiten zu machen. Wir haben folgende Ergebnisse.

Theorem D.2. *Es existiert eine Konstante c, so dass für n groß genug gilt*

$$\mathbb{E}[S_n] \leq n\gamma \leq \mathbb{E}[S_n] + c(n\log n)^{1/2}.$$

Theorem D.3. *Es existiert eine Konstante c, so dass für alle n und alle $\lambda \geq 0$ gilt*

$$0 \leq \Lambda(\lambda) - \frac{1}{n}\Lambda_n(\lambda) \leq \frac{1}{n}\lambda c + \frac{1}{n}\log(2n+1).$$

Die bei beiden Ergebnissen benutzte Zerlegungseigenschaft optimaler globaler Alignments beruht auf einer grafischen Darstellung von Alignments durch Pfade in einem speziellen 2-dimensionalen Gitter. Anhand dieser Darstellung wird auch deutlich, dass optimale Alignments aus mathematischer Sicht sehr eng mit *Perkolationsmodellen* verwandt sind. Im Gebiet der Perkolationsmodelle sind, trotz sehr aktiver Forschungstätigkeiten in den letzten Jahren, noch viele der zentralen Probleme offen.

Feinere Asymptotik

Unserem Ziel, eine feinere Asymptotik für $\mathbb{P}(M_n \geq t)$ angeben zu können, sind wir insofern näher gekommen, als wir zeigen können, dass eine Konstante c existert, sodass gilt

$$\mathbb{P}(M_n \geq t) \leq cn^2t^2 \exp(-\theta^* t).$$

Das Entscheidende bei dieser Abschätzung ist, dass $\exp(-\theta^* t)$ der einzige exponentielle Term in t ist. Ansonsten gibt es auf der rechten Seite nur noch einen Term in t, der polynomial wächst. Wir vermuten, dass die obere Abschätzung *in diesem Sinne* scharf ist, d. h., dass es einen Exponenten η gibt, so dass auch eine untere Abschätzung der Form

$$\mathbb{P}(M_n \geq t) \geq c'n^2t^\eta \exp(-\theta^* t)$$

gilt. Auch wenn wir hierfür keinen rigorosen Beweis liefern können, so haben wir doch einen Ansatz, in dem Theorem D.3 eine entscheidende Rolle spielt.

Praxisrelevanz der Ergebnisse

In der Biologie haben sich mittlerweile einige Software-Werkzeuge etabliert, in denen die Ähnlichkeitssuche mittels optimaler, lokaler Alignments implementiert ist. Dabei wird oft auch ein p-Wert angegeben, um den Anwendern eine bessere Bewertung der gefunden Ergebnisse zu ermöglichen. Hierbei wird allgemein angenommen, dass auch im Fall mit Lücken für zwei Konstanten K^* und θ^*, die natürlich von den zugrunde liegenden Scoringparametern abhängen, gilt

$$\mathbb{P}(M_n \geq t) \sim n^2K^* \exp(-\theta^* t). \tag{8}$$

Diese Vermutung basiert zum Einen auf dem rigorosen Ergebnis im Fall ohne Lücken und zum Anderen auf Simulationsstudien. Von einem Beweis dieser Vermutung ist man allerdings noch um einiges entfernt. Das Hauptergebnis der vorliegenden Arbeit sichert das Analogon von (8) zumindest auf einer gröberen (nämlich der logarithmischen) Skala. In der Tat ist bei der Berechnung des p-Wertes für sehr große t nur der führende Teil der Asympotik interessant. Dies wurde auch in der praxisorientierten Literatur erwähnt.

Darüber hinaus geben unsere Ergebnisse unter der Annahme, dass (8) korrekt ist, eine neue Charakterisierung des Parameters θ^* aufgrund von Theorem D.1. θ^* lässt sich zwar nicht direkt als die eindeutig bestimmte, positive Lösung von

$$\Lambda(\lambda) = 0$$

berechnen, da Λ selbst als Grenzfunktion nur annäherungsweise bestimmt werden kann. Bezeichnet man allerdings für jedes n mit θ_n^* die eindeutig bestimmte, positive Lösung von

$$\Lambda_n(\lambda) = 0,$$

so folgt aus (7)

$$\theta_n^* \searrow \theta^*, \quad n \to \infty.$$

Ausserdem lässt sich in diesem Grenzübergang die Konvergenzrate mit Hilfe von Theorem D.3 abschätzen als

$$\theta_n^* = \theta^* + O\left(\frac{\log n}{n}\right).$$

Chapter 1

Introduction

In this introduction we give a basic overview of the role that alignments play in biology and mathematics. Alignments are an important tool for biologists to get information about the relations between different DNA- or amino acid-sequences. The usage of this tool leads to the mathematical questions on which this thesis is focused. From the mathematical point of view, alignments are best considered to be a special model from the class of percolation models.

1.1 Alignments and biology

1.1.1 Alignments and genealogies

Biologists consider similarities found in two DNA- or amino acid-sequences often to be caused by common descent. Suppose you have two pieces of DNA from which you know that they both stem from the same ancestral sequence. In the process of inheritance (from cell to cell or from organism to organism) the information in the ancestral sequence has been copied over and over again; a process that opens the possibility of a great variety of mutational events. Knowing the ancestral sequence and knowing the complete *genealogy*, i.e. all the changes the sequences underwent during inheritance, would enable us to exactly describe the relationship between the two sequences. However, this information is usually not available to us and what we need is a good guess at how the two sequences are related.

Biologists consider the two main forms of mutational events to be *substitutions* and *insertions/deletions*. Substitution means that in the copying-process some letter at some position is replaced by a different letter. Insertions and deletions clearly refer to the situation, where one or several letters are inserted or deleted during copying.

Considering only these two basic types of mutational events, a guess at how two given sequences might be related should state

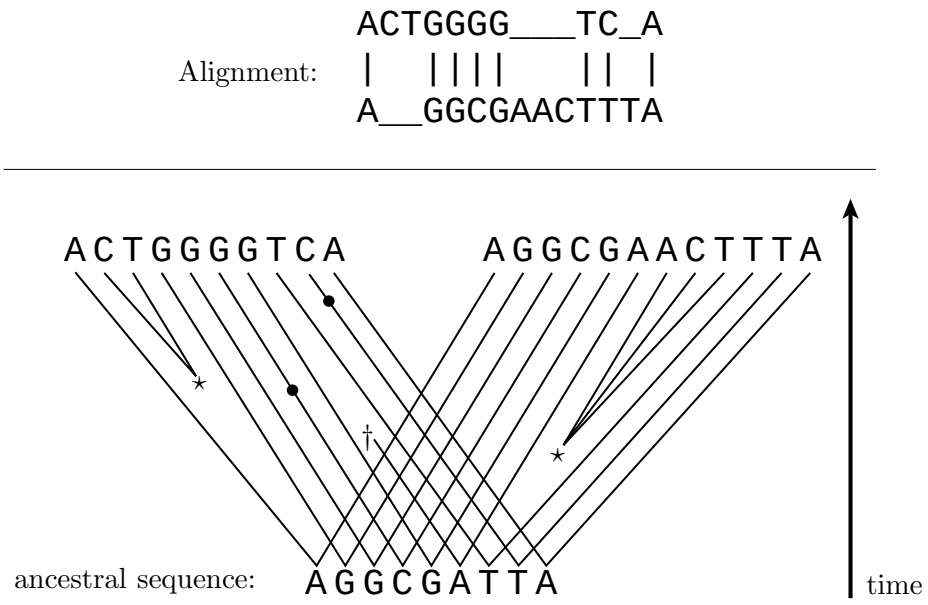

Figure 1.1: *An example for an alignment of two DNA sequences and a genealogy which is presented by this alignment. Substitution events are marked by bullets (•), insertion and deletion events are marked by stars (⋆) and crosses (†), respectively. Observe that this is only one of many possible genealogies that are compatible with this alignment.*

(i) which of the positions in the sequences are *homologous*, i.e. which of the letters stem from the same letter in the ancestral sequence (but might or might not have been changed by some substitutions) and

(ii) where (and which) letters have been inserted or deleted.

This information can conveniently be given using a *gapped alignment*, meaning that one writes down the sequences such that the homologous positions are aligned vertically and gaps in one or the other sequence are used at positions where letters have been inserted or deleted. For an example see Figure 1.1.

Recall that we wanted the alignment to be a *good guess* at the relationship between the two sequences. The value of an alignment can only be judged if we fix some criterion for the alignment's quality. To explain how such a criterion can be chosen in a biologically meaningful way, we have to explain a few more concepts.

1.1.2 The stochastic process of mutations

On the level of the letter sequences the process of evolution is often modelled as a stochastic process. To keep the story simple, substitutions are usually

modelled via a continuous-time Markov-process that acts independently on the different letters. For DNA, the most simple substitution model is the Jukes-Cantor model ([JC69]), where all possible substitutions occur with the same rate. Of course, much more complicated models are conceivable, the next level being one that has different rates for transitions and transversions. It is also possible to model the substitution-process on the amino acid-level. In that case the substitution rates differ a lot between different letter combinations. Still, rate matrices are usually taken to be symmetric, since one wants the stochastic process to be time-reversible.

Also the process of insertions and deletions has to be modelled. The classical model for this has been developed by J. L. Thorne, H. Kishino and J. Felsenstein ([TKF91]). Roughly speaking, after a random time it chooses a random position in the sequence and inserts or deletes a random number of letters at that position. Again, this process is independent at different positions.

With this stochastic process on the level of letter sequences at hand, we can also say something about the quality of alignments. Take two sequences, supposing one of them is the ancestral of the other and the mutation process evolved over some fixed amount of time t. Each of the possible ways the mutation process might have evolved has

(i) a certain probability to occur and

(ii) implies a certain alignment of the two sequences.

So what we get in fact is a probability distribution on the set of all alignments of the two sequences: Just take a given alignment and calculate the probability that the history was such that it produced this very alignment. We can now answer the question about a biologically meaningful good alignment by just taking the Maximum Likelihood alignment, i.e. the one for which this probability is maximized.

A closer look at how the stochastic process of mutations is modelled gives that calculating the probability of an alignment is essentially equivalent to calculating a certain *additive score* of the alignment, which we will shortly explain in more detail. The basic feature here is that this score increases with the alignment's probability. Hence, calculating the Maximum Likelihood alignment is equivalent to calculating the maximum score alignment. Details of this correspondence can be found e.g. in the article of Thorne et al. ([TKF91]) or the book of Durbin et al. ([DEKM98]).

The additive score we have just mentioned is defined as follows. First, there is a summand for each pair of aligned letters in the alignment. The values for the summands are taken from the *scoring matrix*, which has entries for all possible letter pairings. Those entries of course depend on the chosen model for the mutation process and on the time t. They are usually positive for pairs of equal letters and negative for pairs of different letters. The most

classical family of such scoring matrices are the PAM-matrices (PAM for accepted point mutations). They have been derived from a probabilistic model that has been adapted to a certain set of given aligned sequences. For details cf. [DSO78].

Second, for each gap in the alignment there is a penalty subtracted. The calculation of this penalty is based on two nonnegative parameters Δ and δ. For each gap subtract Δ once and subtract δ multiplied by the length of the gap. Thus the gap-penalty is an affine linear function in the gap's length. The parameters Δ and δ are called *gap-open* and *gap-extension penalty*, respectively. Again, the values of these two parameters depend on the underlying stochastic process and on the time t.

As we have already mentioned, determining the maximum score alignment is equivalent to determining the Maximum Likelihood alignment. At that stage it is not yet clear if, for any two given sequences, this maximization can be done in reasonable time. Fortunately, the situation is such that this is the case. Using Dynamic Programming techniques, an algorithm can be implemented which finds the maximal score alignment of two given sequences in time proportional to the product of the sequence lengths. Dynamic Programming techniques are especially convenient here, since the mutation process can be interpreted as a certain Hidden Markov model, which has the two dependent letter sequences as its output. Details about Hidden Markov models, Dynamic Programming and, especially, alignment algorithms can be found in [DEKM98]. In the biological world the algorithm that determines the optimal (global) alignment of two given sequences, is known as the *Needleman-Wunsch algorithm* (cf. [NW70]).

1.1.3 Getting rid of the ancestral sequence

We have just described how an optimal score alignment can be interpreted as a Maximum Likelihood alignment under some stochastic model of evolution, when one of the sequences is the ancestral and the second sequence has evolved from it. However, usually, when we are given two related sequences, it is not the case that one of them is the ancestral of the other, but both stem from some ancestral sequence that is not available.

In order to circumvent this problem, the stochastic process of mutations is taken to be *time-reversible*. This roughly means that if we are given two *typical* letter sequences and we know that one of them is the ancestral of the other, we can not say *which* of them is the ancestral one. This is also reflected by the fact that the additive scoring system we have just described, does not depend on which sequence is on top and which one is on the bottom of the alignment. Here the term *typical* refers to sequences that one observes after the process of mutations has acted for a very long time on some starting sequence.

In our situation this time-reversibility means that we can change the time

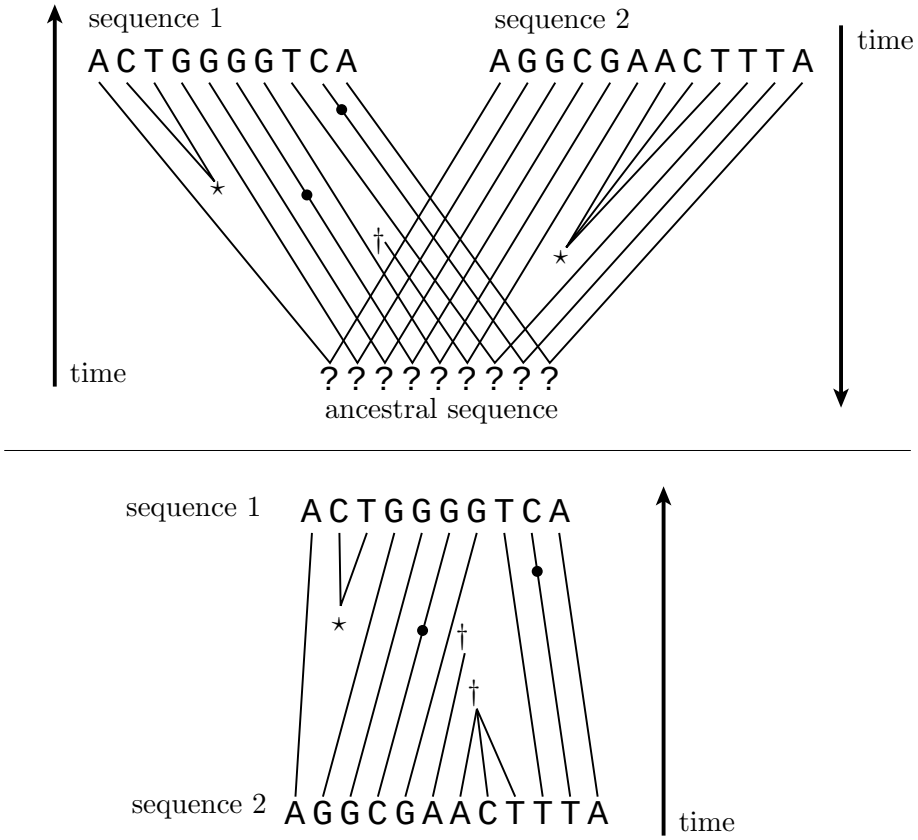

Figure 1.2: *Since the process of mutations is time-reversible, one can always assume that one of the two sequences is the ancestral of the other.*

direction at one branch of our geneology in Figure 1.1 and therefore always assume that for two given sequences one is the ancestral of the other. This principle is shown in Figure 1.2. The mathematical details can be found in [TKF91].

1.1.4 Local similarities

Up to now we have only described the situation, where one is interested in an optimal alignment that relates *the whole* of the two given sequences. This kind of optimal alignment is therefore usually called optimal *global* alignment. However, it is often the case that one does not know whether there is a relationship between all of the sequences. In this case it makes more sense to search for an optimal alignment of two *optimally chosen subintervals* of the sequences. One speaks then of an optimal *local* alignment. On the score side, this difference is only expressed in the way gaps at the two very ends of the alignment are treated. If it is allowed to have arbitrar-

ily long gaps at one or both of the ends *at no cost*, then the optimization leads to the optimal local alignment. The corresponding algorithm is known as Smith-Waterman algorithm (cf. [SW81]). Again it is based on Dynamic Programming techniques and has a run-time which is proportional to the product of the sequences lengths.

1.1.5 Alignments in biological practice

In biological practice alignments are used in two main ways.

The first concerns the detection of similarities between a newly determined sequence and a database of known sequences. In the past years the techniques for determining new DNA or amino acid sequences have undergone an unexpected development and acceleration. It is simply due to the quantity of newly determined sequences that it is impossible to examine them one by one. Automatized mechanisms are needed to get a first idea of whether a sequence contains some interesting parts.

Searching for extraordinarily good alignments across a database of known sequences is such a mechanism. The most popular being BLAST (from: basic local alignment search tool, cf. [AGM$^+$90]) and FASTA (cf. [PL88]). In principle, such tools compare the query sequence with each sequence in the database, every time determining an optimized alignment together with its score. Then the best of those alignments are given back. Searching for *local* alignments makes sense, since the sequencing does not necessarily start or stop at biologically meaningful positions on the molecule. BLAST and FASTA both use the additive scoring scheme just described, but they do *not* use an exact Smith-Waterman algorithm to determine their optimized alignments. Instead, they use faster algorithms for performance reasons. An implementation of the exact Smith-Waterman algorithm is available in the usual FASTA packages under the name SSEARCH.

BLAST and FASTA are available for both DNA and amino acid sequences. For the 4-letter DNA alphabet the required scoring matrices are relatively small. The simplest matrices distinguish only between matches and mismatches. These matrices correspond to the Jukes-Cantor model of substitutions. For the 20-letter amino acid alphabet, the most popular scoring matrices are the so-called PAM and the BLOSUM ([HH92]) matrices — they can be used in both packages.

The choice of the gap penalties is another crucial issue. Choosing too small gap penalties leads to scoring schemes which incorporate into the optimal alignment as many matching letter pairs as possible, the gap sizes between those matching letter positions playing only a minor role. Such scoring schemes would be unable to detect local similarities between the sequences, since scores increase with the number of letters used in the alignment. It is therefore necessary to choose high enough gap penalties. On the other hand, if gap penalties are chosen too large, one prevents the occurence

of gaps in optimal alignments. This is also not desirable, especially if one wants to detect distant relationships between sequences. We see that the scoring scheme has to be chosen carefully in order to have the necessary sensitivity and produce the desired results in the database search.

Optimal *global* alignments are central for the second way in which alignments are used. As already mentioned above, alignments can be interpreted as a statement on the details of the relationship between two sequences. Those details might, e.g., be needed to construct phylogenies. A good global alignment of two sequences is needed, since unrelated parts of the sequences can be discarded beforehand. Again, the scoring scheme has to be chosen carefully regarding the biological needs.

In this thesis we are mainly concerned with problems arising from the first use of alignments just described. Especially, we are interested in the statistical problems related to it.

1.1.6 The main problem: statistical significance

The main problem, when searching for similarities between a new sequence and a large collection of sequences in a database, is how to evaluate the result. The output of the alignment algorithm consists of an optimal local alignment together with its score. Somehow, the score quantifies the degree of similarity that can be found in the alignment. But how high does a score have to be in order to indicate a similarity that is worth looking at? This question can be best answered by calculating the *statistical significance* of the score value.

Statistical significance means: Suppose we would build up the entire database with independent random sequences and would also take an independent random query sequence. What is then the probability of getting an optimal alignment score at least as high as the value in question? If the probability is small, then this might indicate a degree of similarity that is unlikely to occur only by chance.

We have already mentioned in the Preface that we use independent random sequences to give sense to the notion of *unrelated sequences*. For a statistical evaluation of alignments scores it is necessary to agree on a stochastic model to generate the random sequences. This model then takes the usual role of the *null model* in statistics. In accordance with widespread biological practice, we use the simplest model where all letters of the random sequence are drawn independently according to some distribution on the underlying alphabet. The letter-weights of this distribution are chosen according to letter-frequencies observed in real sequences (usually according to [RR91]).

The main problem with this simple model is that there is some statistical evidence that real-world sequences do in fact not look like the sequences produced by this model. Especially for DNA sequences, there are strong correlations between every third letter, since the reading-frame for trans-

lation into the amino acid alphabet is three letters long. Therefore, other — more refined — models might be more appropriate for the generation of unrelated sequences. Nevertheless, from the mathematical point of view, already the simple null model is difficult enough, so we leave the problem of dealing with more complicatd null models for further investigations.

From the practitioner's point of view one can still try to use simulations, if one is not able to characterize the optimal score's distribution theoretically. But observe that there are two issues. First, since we are interested in the extreme upper tail of the distribution, it is clear that a naive Monte Carlo-simulation would need a very large number of samples and therefore be very time-consuming. Second, one would like to know the distribution for different alphabets, different sequence-compositions and, most of all, different scoring schemes. Ideally, a practitioner should be allowed to choose any of these ingredients as he wants to. It would thus be desirable to be able to characterize the optimal score's distribution without doing time-consuming simulations for each set of parameters. Up to now this problem has been circumvented in the existing implementations by only allowing a restricted set of base compositions, scoring matrices and gap penalties.

1.1.7 Many sequences vs. two sequences

Although the final aim is to evaluate the optimal score produced by a comparison of one sequence with all the sequences from a large database, this can easily be reduced to the problem of comparing two sequences. We agreed that under the null model all the sequences involved are independent, thus each comparison of the query sequence with a sequence from the database produces an independent optimal score. The overall optimal score of the database comparison is then simply the maximum of these independent scores. The statistics behind maxima of independent random variables is quite well understood, a classical reference being the book of Embrechts, Klüppelberg and Mikosch ([EKM97]). The determining quantity is the asymptotic behaviour of the upper (since we are doing a *maximization*) tail of the single score's distribution. In the case of optimal local alignments without gaps this has also been described by R. Spang and M. Vingron in [SV98]. There it is also discussed which problems arise from working with real databases.

So we can finally formulate the proper statistical problem we address in this thesis. Consider the optimal local alignment score one obtains when comparing two sequences of independent random letters. How does the tail of this score's distribution behave asymptotically, when both the score and the sequence lengths tend to infinity?

This question basically has two parts.

(i) There should be some functional relationship between the score and the sequence lengths in the asymptotic approximation. This relation-

ship should be universal in the sense that it should not depend on input parameters such as letter distribution, scoring scheme, etc.

(ii) On the other hand there will be some quantities in the asymptotic approximation, which depend on those input parameters. The exact form of this dependence has to be determined in order to give useful results for practice.

1.2 Optimal alignments and percolation theory

We start by describing the relationship between optimal alignments and percolation models. When looked at from the formal point of view, optimal alignments are closely related to classical models from percolation theory, especially to first-passage percolation. The classic questions from percolation theory appear also to be central to optimal alignments. Anyhow, new features and problems, unknown to classical percolation theory, appear in our setting. These new aspects derive from the concept of optimal local scores. In classical percolation models there is no equivalent to this concept. The questions arising from it are merely related to certain aspects of the large deviations theory of random walks. It is thus the combination of these two fields that justifies from the mathematical point of view why one should deal with it. We will now describe these issues in more detail.

1.2.1 Classical first-passage percolation

First-passage percolation has been introduced 1965 by Hammersley and Welsh in [HW65]. In fact it was the initial spark for the development of the theory of subadditive processes. The main surveys of the subject are still Kesten's Saint Flour lecture note [Kes84] and Smythe and Wierman's book [SW78]. We start by a short description of the model and the basic results.

We restrict ourselves to the two-dimensional case. Consider the lattice with vertex set $\mathbb{Z}^2$ and with edges between all pairs of neighbouring vertices. Assign to each edge e a random variable T_e, where the T_e's should be non-negative and drawn independently from the same distribution. A path π in the lattice is a sequence of neighbouring edges that connects one vertex with some other. To each path a random variable T_π is assigned by summing up the T_e's along the path.

The term *percolation* refers to the following interpretation. Think of water percolating through a porous medium. The geometric structure of the medium is given by the lattice, but the medium is very inhomogeneous and this is modelled by the random variables. Interpret T_e as the time the water takes to travel through edge e. Then T_π is the time it takes to travel through the path π. If the water is supplied at the origin, how long

does it take until it reaches a certain point $\mathbf{x}$? To answer this question, minimize T_π over all paths that connect the origin with the point $\mathbf{x}$ and denote the resulting quantity by $S_{0,\mathbf{x}}$. Percolation theory is concerned with the behaviour of this random variable, or variants of it, in the asymptotics where $\mathbf{x}$ is far away from the origin.

To explain some of the results, denote by C_t the set of lattice points, for which $S_{0,\mathbf{x}} \leq t$. C_t is a randomly-growing connected cluster in $\mathbb{Z}^2$ which contains the origin. It follows from subadditivity arguments that asymptotically this cluster grows linearly with a deterministic shape $B \subseteq \mathbb{R}^2$, meaning that

$$\frac{C_t}{t} \xrightarrow[t\to\infty]{} B \quad \text{a.s.}$$

For some unit vector $\mathbf{e} \in \mathbb{R}^2$, the value

$$\gamma_{\mathbf{e}} := \sup\{\lambda > 0 \colon \lambda \cdot \mathbf{e} \in B\}$$

is usually called *time constant of direction* $\mathbf{e}$. The shape B is known to be convex, but its proper form depends on the distribution of the T_e's. There is no general result about this dependence of B on the underlying distribution, not even for special directions of growth.

Even harder is the question of the fluctuations of C_t around tB. It is conjectured (cf. [NP95] and the references therein) that they are of the order of $t^{1/3}$, but there has been no rigorous proof up to now. It is interesting that such a behaviour of the fluctuations is conjectured for a much broader class of growth models. We will not go deeper into this, but merely refer to the survey article of [KS91], which gives a good introduction into the field.

Much of what we have just stated holds also for the special model of *directed* first-passage percolation. Here, the set of paths, over which one is allowed to minimize, is restricted to *increasing* paths, where increasing means that the vertices visited by the path have to increase in the usual partial ordering on $\mathbb{Z}^2$. We mention this as we will see in the next subsection that optimal alignments are very closely related to directed percolation.

1.2.2 Optimal alignments as a percolation model

There is a way of graphically representing alignments of two sequences that immediately reveals the connections to percolation theory. We will explain it here in an informal way, precise definitions can be found in Section 4.1. This correspondence has been known at least since Arratia and Waterman's paper [AW94].

Consider Figure 1.3. There, the alignment from Figure 1.1 is represented by a path in some special lattice. In fact, any possible alignment of the two sequences can be drawn as some increasing path into this grid. Writing the sequences as shown to the sides of the grid's rectangle, gives a one-to-one correspondence between aligned positions in the alignment and a collection

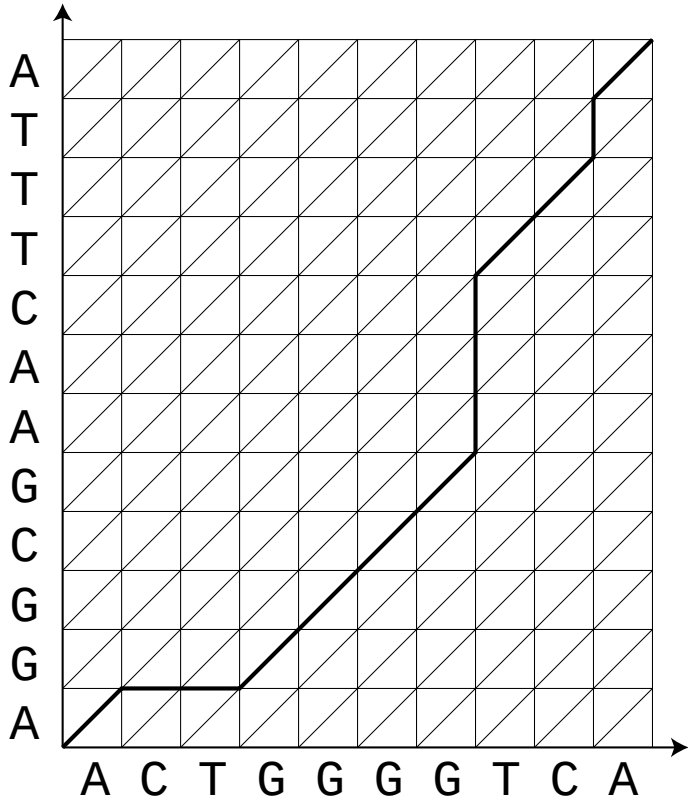

Figure 1.3: *Representing the alignment from Figure 1.1 by an increasing path in a special lattice. Diagonal edges in the path correspond to aligned positions in the sequences.*

of diagonal edges in the grid that can be connected by horizontal and vertical edges to result in an increasing path. These horizontal and vertical edges correspond to the unaligned positions.

Let us now see that the score of an alignment can be calculated in a way that strongly reminds of first-passage percolation theory. First we assign a value to each edge in the lattice. For a diagonal edge, read off the two letters belonging to it and take the corresponding entry from the scoring matrix. To the horizontal and vertical edges, just assign the negative of the gap-extension penalty. Observe that randomizing the sequences leads to a random assignment of values to the diagonal edges. An important difference to classical first-passage percolation is that these values are no longer independent, but carry a special dependence structure.

The score of the alignment can now be calculated as follows. We start by summing up the edge-values along the path that represents the alignment. Were the gap-open penalty equal to zero, we would be done. But in cases where the gap-open penalty is unequal to zero we have to subtract it once

for each of the horizontal and vertical stretches in the alignment path. This way of assigning a value to each increasing path in the lattice is quite similar to first-passage percolation. The extra part, which has its origins in the gap-open penalty can be interpreted as a penalty for the path's *roughness*.

Optimal alignments can now be determined in much the same way as in first-passage percolation (the only difference being the use of a maximization instead of a minimization; thus it would be more appropriate to speak of *last* passage percolation). For the optimal *global* alignment, maximize the score over all increasing paths that start in the lower left corner and end in the upper left corner of the grid. For the optimal *local* score, maximize over all increasing paths that start and end at any two vertices.

As we have already mentioned above, we want the randomness in this picture to be generated by two independent random letter sequences. So let us suppose that $\mathbf{X} = (X_1, X_2, \ldots)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots)$ are two sequences of letters from the underlying alphabet $\mathcal{A}$, in which all letters have been drawn independently according to a given distribution μ on $\mathcal{A}$. Writing these two infinite sequences along the axis as in Figure 1.3, we get a percolation model on the whole of the lattice, in which the diagonal values are randomized. If we denote the scoring matrix by K, then the edge that connects the vertices $(i-1, j-1)$ and (i, j) in the lattice, gets the value $K(X_i, Y_j)$. Observe that for the square $[0, n]^2$ in the lattice, we have a total of n^2 diagonal edges to which values are assigned but only $2n$ independent random letters which generate this random environment. It is clear that this produces certain dependences between the diagonal values.

The two basic quantities needed in the following are the optimal *global* and the optimal *local* score of two length n sequences. We will denote them by $S_{0,n}$ and $M_{0,n}$, respectively. It is clear that $S_{0,n}$ is the score of the optimal path which connects $(0, 0)$ with (n, n), whereas $M_{0,n}$ is the optimal score among all paths that start and end in the square $[0, n]^2$.

1.2.3 Questions arising from the model of optimal scores

The basic results for $S_{0,n}$ can, as in classical percolation theory, be derived using superadditivity. Therefore again we have that $S_{0,n}$ grows asymptotically linearly in n, meaning that

$$\lim_{n\to\infty} \frac{S_{0,n}}{n} = \gamma \quad \text{a.s.},$$

where

$$\gamma := \sup_{n>0} \frac{\mathbb{E}[S_{0,n}]}{n} = \lim_{n\to\infty} \frac{\mathbb{E}[S_{0,n}]}{n}$$

(the last equation being an existence statement for the limit). Now γ should rather be called *growth constant* instead of *time constant*, since the values we assigned to the edges can be negative.

In fact, γ can even be negative. When this is the case, the average optimal global score of two sequences deteriorates when the sequences get longer. This turns out to be the interesting case when we are searching for local similarities. It prevents us from just taking the longest alignment in order to accumulate as much of the score as we can. Instead, if there are local parts in the sequences which share some similarity, an optimal alignment of those subparts gets a positive score. The best of those similarity-sharing subparts are then found by the optimal local alignment algorithm.

Therefore, in the case of a negative γ, the distribution of the optimal local score will be of a completely different character than that of the optimal global score. Loosely, we could say that we are looking for good local parts in a bad global environment. It is in this regime, where we want to characterize the distribution of the optimal local score of two random letter sequences.

The fact that the optimal local score behaves completely different depending on whether γ is positive or negative, has first been observed by Arratia and Waterman in [AW94]. They call this phenomenon a *phase transition* for the optimal local score. The two phases described are the *global regime*, where $\gamma > 0$ and the optimal local score behaves basically as the optimal global score, and the *local regime*, where $\gamma < 0$. Since the main results of this thesis can be used to simplify the original proof in [AW94], we will come back to this later in Section 4.7. We are primarily interested in the local regime, and many of the central ideas of this thesis only work there. Therefore, we assume from now on that we are in the local regime, unless otherwise stated.

Coming back to classical percolation theory, observe that there are no quantities playing the same role as optimal local scores. Hence, we have to look at percolation models in an essentially new way. As we have already mentioned at the beginning of this section, the questions we address here are merely related to the theory of large deviations of random walks. We will explain this in more detail in Section 2.1.

Chapter 2

Basic ideas and concepts

In this chapter we will explain some of the ideas and concepts behind the results of this thesis. The aim is to give an easily readable yet concise introduction to the results proved rigorously in the later chapters. Here we merely use heuristic calculations, for which we do not claim to fulfil the requirements of mathematical precision. Nevertheless, everything we state is proved in the literature or will be proven rigorously later in this thesis. We always give the relevant references.

2.1 The level of large deviations

Characterizing the distribution of optimal local scores is the main task of this thesis. More precisely, we are interested in the asymptotic behaviour of its tail. Several levels of exactness are conceivable, when characterizing the tail of a distribution. The first of these is certainly the level of large deviations, where one deals only with questions related to the leading term of the tail's decay. Often this leading term is exponentially decaying and the task is to characterize the *order* of this exponential decay and its *rate*.

We will explain our answers to both questions for the case of optimal local scores in this section. For methodological reasons we start with the gapless case, where in fact much more is known about the optimal local score's distribution, but the restriction to a large deviations level already reveals the central concepts.

2.1.1 The gapless case

Considering gapless alignments simply means that in the optimization, which is carried out in the calculation of the optimal local or global scores, one is not allowed to consider alignments with gaps. We start by examining this much easier gapless case, since on the large deviations level the central ideas also carry through to the general case. Recall that we still assume that

$\gamma < 0$.

The complete picture in the gapless case has been given by Amir Dembo, Samuel Karlin and Ofer Zeitouni ([DKZ94a] and [DKZ94b]). They give a distributional limit theorem by proving that

$$\lim_{n\to\infty} \mathbb{P}\left(M_{0,n} - \frac{2\log n}{\theta^*} \leq t \right) = \exp(-K^* \exp(-\theta^* t)), \tag{2.1}$$

where θ^* and K^* are two positive constants. We will now explain this result, and especially the role of the constant θ^*, in more detail.

Let us start with the form of the limiting distribution. Distribution functions of this double-exponential form are known as *extreme value distributions of type I* (or *Gumbel distributions*) and usually appear in the theory of extreme values (cf. the book of Embrechts, Klüppelberg and Mikosch, [EKM97]). For example, if one considers a sequence $(X_i)_{i\geq 0}$ of independent Exp(1)-distributed random variables the following limit holds

$$\lim_{n\to\infty} \mathbb{P}(\max(X_1, \ldots, X_n) - \log n \leq t) = \exp(-\exp(-t)).$$

Crucial for this result is, apart from the independence, the behaviour of the tail of the distribution. Similar limit theorems hold for all distributions with exponentially decaying tails. In fact, an easy substitution shows that the maximum of n^2 independent and identically distributed random variables, with an upper tail that decays like $K^* \exp(-\theta^* t)$, behave in the limit exactly as $M_{0,n}$ in equation (2.1). Is it the case that in return $M_{0,n}$ can be interpreted as such a maximum? We will argue next that the answer is positive.

To start our investigation, we have to introduce another optimal score. Let $\mathbf{i}$ be some vertex in the lattice shown in Figure 1.3 on page 11 and define $T_{\mathbf{i}}$ as the maximal score one can get, when maximizing over all paths that start at vertex $\mathbf{i}$. The key lies in the observation that

$$M_{0,n} \cong \max_{\mathbf{i}\in[0,n]^2} T_{\mathbf{i}}, \tag{2.2}$$

where we do not have an exact equation because of edge effects.

We will now start by arguing that the upper tail of the $T_{\mathbf{i}}$'s, which all have the same distribution, has the right exponential decay, i.e. that we have

$$\mathbb{P}(T_{\mathbf{i}} > t) \simeq \exp(-\theta^* t) \tag{2.3}$$

(for the definition of different asymptotic relations cf. Appendix A.1). This will then help us understand that the approximation in (2.2) is not so bad for large t, since the relevant $T_{\mathbf{i}}$'s do not use too long paths in their optimization.

Let us analyze $T := T_{(0,0)}$. In the percolation picture, the restriction to gapless alignments means that we are restricted to paths that stay on a single diagonal. This makes it possible to characterize T as

$$T := \max_{k\geq 0} S_{0,k},$$

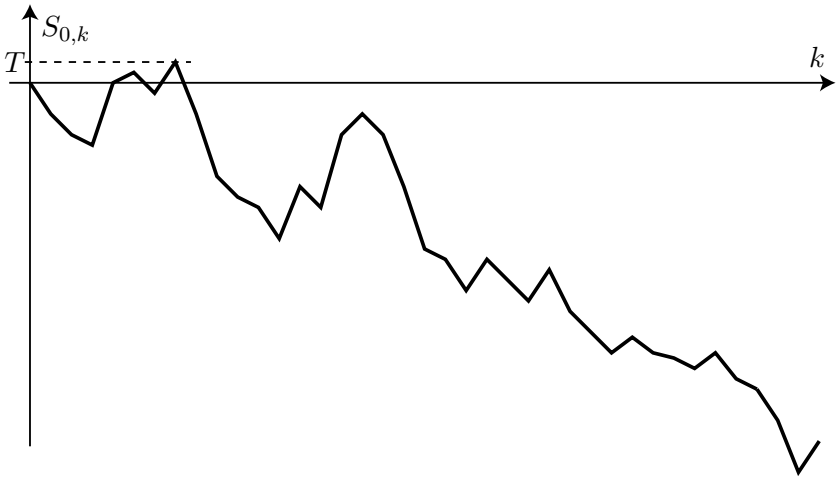

Figure 2.1: *The sequence* $(S_{0,k})_{k\geq 0}$ *forms a random walk with negative drift but some positive steps.*

where of course we set $S_{0,0} = 0$. Also the $S_{0,k}$'s can be nicely characterized: there remains only one possible path in the maximization. This implies that we can write

$$S_{0,k} = \sum_{l=1}^{k} K(X_l, Y_l),$$

showing that it is nothing but the sum of i. i. d. random variables. The condition $\gamma < 0$ reduces thus to the requirement that

$$\gamma = \mathbb{E}[K(X_1, Y_1)] < 0.$$

Putting this together, we have that the process $(S_{0,k})_{k\geq 0}$ is a random walk with negative drift but with a positive probability of taking positive steps, since usually scoring matrices have positive entries on their diagonal. For a picture of such a process see Figure 2.1. Also, T is nothing but the overall maximum of such a random walk.

To characterize the rate of decay θ^* in (2.3) we first have to introduce the *logarithmic moment generating function* Λ of the score of a random letter pair by

$$\Lambda(\lambda) := \log \mathbb{E}[\exp(\lambda K(X_1, Y_1))] \tag{2.4}$$

which is convex and has $\Lambda'(0) = \gamma < 0$. Then θ^* is given by the unique positive zero of Λ, i.e.

$$\Lambda(\theta^*) = 0.$$

To understand why this characterization is valid we first have to examine the role of Λ in more detail. In the following we will give a heuristic argument, which will also be helpful in the gapped case.

It is a basic fact from the large deviations theory of sums of i.i.d random variables that the (large deviations) rate function

$$r(q) := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(S_{0,n} \geq qn)$$

for sums of i. i. d. random variables, where each of the summands is distributed as $K(X_1, Y_1)$, is given by the *Fenchel-Legendre transform* (or *convex conjugate*, cf. Appendix A.2) Λ^* of Λ, i.e. by

$$\Lambda^*(q) := \sup_{\lambda\in\mathbb{R}}\{\lambda q - \Lambda(\lambda)\} \tag{2.5}$$

(we refer to the [DZ93] for more precise information about large deviations theory). This can be easily seen with the help of the following heuristic calculation. Observe first that the supremum in (2.5) is attained at the value λ_q for which $\Lambda'(\lambda_q) = q$. We can embed the distribution $\mathbb{P}_0 := \mu \times \mu$ on $\mathcal{A} \times \mathcal{A}$ into an exponential family of distributions on $\mathcal{A} \times \mathcal{A}$ by defining

$$\frac{\mathrm{d}\mathbb{P}_\lambda}{\mathrm{d}\mathbb{P}_0}(x, y) := \exp(\lambda K(x, y) - \Lambda(\lambda)).$$

Of course this extends in a natural way to an exponential family of distributions on pairs of letter sequences of length n, i.e. on $(\mathcal{A} \times \mathcal{A})^n$. This way of switching between equivalent measures using exponential weights is also sometimes referred to as *exponential tilting*.

For a given $q > \gamma$ we now want to calculate the probability $\mathbb{P}_0(S_{0,n} > nq)$ (an analogous argument is valid for $q < \gamma$ and $\mathbb{P}_0(S_{0,n} < nq)$). Observe that

$$\mathbb{E}_\lambda[K(X_1, Y_1)] = \Lambda'(\lambda)$$

so that under $\mathbb{P}_{\lambda_q}$ we have $\mathbb{E}_{\lambda_q}[S_{0,n}] = nq$ and an easy calculation gives

$$\begin{aligned}\mathbb{P}_0(S_{0,n} > nq) &= \mathbb{E}_0[1; S_{0,n} > nq] \\ &= e^{-n(\lambda_q q - \Lambda(\lambda_q))}\mathbb{E}_{\lambda_q}[\exp(-\lambda_q(S_{0,n} - nq)); S_{0,n} > nq] \\ &= e^{-n\Lambda^*(q)}\mathbb{E}_{\lambda_q}[\exp(-\lambda_q(S_{0,n} - nq)); S_{0,n} > nq].\end{aligned} \tag{2.6}$$

It is well known that the *truncated moment generating function*

$$\mathbb{E}_{\lambda_q}[\exp(-\lambda_q(S_{0,n} - nq)); S_{0,n} > nq]$$

is $\Theta(n^{-1/2})$ (cf. [BR60] or [Hir98]; see Appendix A.1 for notation), so

$$r(q) = \Lambda^*(q).$$

Let us now turn to local alignments. Using a Laplace-type approximation one sees easily that

$$\mathbb{P}_0(T > t) = \mathbb{P}_0(\max_{i\geq 0} S_{0,i} > t) \simeq \max_{i\geq 0} \mathbb{P}_0(S_{0,i} > t).$$

The analysis of the large deviations of $S_{0,i}$ now gives (with $q = t/i$)

$$\begin{aligned} \mathbb{P}_0(S_{0,i} > t) &= \mathbb{P}_0(S_{0,i} > qi) \\ &\simeq e^{-i\Lambda^*(q)} = e^{-i(\lambda_q q - \Lambda(\lambda_q))} \\ &= e^{-t\left(\lambda_q - \frac{\Lambda(\lambda_q)}{\Lambda'(\lambda_q)}\right)}, \end{aligned} \tag{2.7}$$

where the last equality follows from $\Lambda'(\lambda_q) = q$. Recall that q still depends on i (although we supressed this dependence in the notation) so we have

$$\lim_{t\to\infty}\left(-\frac{1}{t}\log\max_{i\geq 1}\mathbb{P}_0(S_{0,i} > t)\right) = \inf_{\lambda:\ \Lambda'(\lambda)>0}\left\{\lambda - \frac{\Lambda(\lambda)}{\Lambda'(\lambda)}\right\} = \theta^*.$$

As a corollary one could state that

$$\max_{i\geq 0}\mathbb{P}_0(S_{0,i} > t) \simeq \mathbb{P}_0(S_{i^*} > t) \simeq e^{-t\theta^*},$$

where $i^* = t/\Lambda'(\theta^*)$.

This has the following interpretation. Under $\mathbb{P}_0$ each summand in $S_{0,i}$ has negative expectation. If we want to know how improbable it is (on an exponential scale) to attain a large value t we can calculate how much it costs to tilt the distribution $\mathbb{P}_0$ such that under the tilted distribution t is the typical value. The tilting cost depends on two things. First we have to pay for each letter pair we want to tilt and second the cost increases the more we tilt. If we choose i too small, we do not have to tilt too many letter pairs, but each of them has to be tilted very much in order to have an overall expectation of t. Whereas if we choose i too large we have to tilt too many letters although each of them only has to be tilted a little bit. The value i^* represents the right balance between this two sides. It is the number of letters one should choose if one wants to optimize the probability of attaining score t.

Let us now discuss in some detail the asymptotics behind (2.2). Fix a large t and an even larger value of n. The large deviations arguments just given, state that if we have a starting point $\mathbf{i}$ for which $T_{\mathbf{i}}$ exceeds the value of t, then the piece of the diagonal that starts at point $\mathbf{i}$ and realizes this score has approximately length i^*. So if we restict in (2.2) the maximization to starting points $\mathbf{i} \in [0, n - i^*]^2$, all optimizing paths should begin and end in $[0, n]^2$.

For our database search application we are mainly interested in a reasonable asymptotics for $\mathbb{P}(M_{0,n} \geq t)$. It is clear that we should let go t and n together to infinity. On one hand n should grow faster than t, since an optimal local alignment that gives a score of t is approximately of length $i_t^* := t/\Lambda'(\theta^*)$. Taking n too small would simply not let enough room for such an alignment and therefore make it impossible to occur. On the other hand it is clear that

$$\mathbb{P}(\max_{\mathbf{i}\in[0,n]^2} T_{\mathbf{i}} > t) \simeq \mathbb{P}(T > t) \simeq e^{-t\theta^*}$$

only holds if n^2 does *not* grow exponentially in t.

Putting this together we can state that when n grows faster than t, but not exponentially fast in t, we have

$$\mathbb{P}(M_{0,n} > t) \simeq \mathbb{P}(\max_{\mathbf{i}\in[0,n]^2} T_{\mathbf{i}} > 0) \simeq e^{-t\theta^*}.$$

Of course much finer asymptotics for this tail have been given, to which we will come back in Subsection 5.1.1.

2.1.2 The gapped case

The heuristic argumentation we have just given for the gapless case can also be transferred to the case where gapped alignments are allowed. In fact, proving these results rigorously, is one of the main results of this thesis. This will be done in Sections 4.5 and 4.6.

Always keep in mind that we are working in the local regime, i.e. that we have

$$\gamma < 0.$$

The problem in the gapped case is that $S_{0,n}$ is not a simple sum of n i. i. d. random variables as in the gapless case. A complicated maximization over many dependent paths is involved in the determination of its value. Nevertheless, we can make quite similar calculations as in the gapless case.

To do this define for each n the logarithmic moment generating function Λ_n of $S_{0,n}$ as

$$\Lambda_n(\lambda) := \log \mathbb{E}[\exp(\lambda S_{0,n})].$$

Again, we can use exponential tilting and define for each $\lambda > 0$ and each n a measure $\mathbb{P}_{\lambda,n}$ on $\mathcal{A}^n \times \mathcal{A}^n$ via

$$\frac{\mathrm{d}\mathbb{P}_{\lambda,n}}{\mathrm{d}\mathbb{P}_0}(\mathbf{x},\mathbf{y}) := \exp(\lambda s(\mathbf{x},\mathbf{y}) - \Lambda_n(\lambda)),$$

where we denote by $s(\mathbf{x},\mathbf{y})$ the optimal global score of the two fixed sequences $\mathbf{x}$ and $\mathbf{y}$. For a given n and $q > \gamma$ choose $\lambda_{n,q}$ such that under $\mathbb{P}_{\lambda_{n,q}}$ the expectation of $S_{0,n}$ is nq, i.e.

$$\mathbb{E}_{\lambda_{n,q},n}[S_{0,n}] = \Lambda_n'(\lambda_{n,q}) = nq.$$

Essentially the same calculation as in (2.6) now gives

$$\begin{aligned}\mathbb{P}_0(S_{0,n} > nq) &= \mathbb{E}_0[1; S_{0,n} > nq] \\ &= e^{-n[\lambda_{n,q}q - (1/n)\Lambda_n(\lambda_{n,q})]}\mathbb{E}_{\lambda_{n,q},n}[\exp(-\lambda_{n,q}(S_{0,n} - nq)); S_{0,n} > nq]. \quad (2.8)\end{aligned}$$

Compare the crucial exponent here, which is

$$\lambda_{n,q}q - (1/n)\Lambda_n(\lambda_{n,q}), \quad (2.9)$$

with the corresponding exponent in (2.6). In (2.6) only the convex function Λ appeared instead of $(1/n)\Lambda_n$ and also $\lambda_{n,q}$ did not depend on n. Although in the gapped case the situation is more complicated, we still have that there exists a convex function Λ with $\Lambda'(0) = \gamma < 0$ and

$$\Lambda(\lambda) = \sup_{n>0} \frac{1}{n}\Lambda_n(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(\lambda) \tag{2.10}$$

and that

$$\lambda_{n,q} \xrightarrow[n\to\infty]{} \lambda_q.$$

Hence asymptotically (2.9) can again be interpreted as the Fenchel-Legendre transform Λ^* of Λ evaluated at the value of q, since we have

$$\Lambda'(\lambda_q-) \le q \le \Lambda'(\lambda_q+),$$

which implies that

$$\Lambda^*(q) := \sup_{\lambda}\{\lambda q - \Lambda(\lambda)\} = \lambda_q q - \Lambda(\lambda_q).$$

Furthermore, it is heuristically plausible that, as in the gapless case, the expectation in (2.8) does not contribute to the exponential rate of decay of $\mathbb{P}_0(S_{0,n} > nq)$. All this gives that

$$\mathbb{P}_0(S_{0,n} > nq) \simeq e^{-n\Lambda^*(q)},$$

meaning that Λ^* is the rate function of the upper tail of $S_{0,n}$. That this is the case for general independent superadditive processes has first been observed by Hammersley in [Ham74]. We will state it more precisely in Section 3.2.

To derive the corresponding large deviations result for local scores is, at least in this heuristical argumentation, straightforward. Again, if we define the $T_{\mathbf{i}}$'s and T as in the gapless case, we roughly get that

$$\mathbb{P}_0(T > t) \simeq \mathbb{P}_0(\max_{i\ge 1} S_{0,i} > t) \simeq \max_{i\ge 1} \mathbb{P}_0(S_{0,i} > t).$$

Of course, in the maximization we should consider all paths that start in the point $(0,0)$ and not only those that end somewhere on the main diagonal. We will respect this issue in our rigorous proof of the result in Section 4.5, whereas here we go on with the simplified version.

Exactly as in (2.7), we can deduce that (with $q = t/i$, λ_q such that $\Lambda'(\lambda_q) = q$ and where we assume for simplicity that Λ is differentiable)

$$\begin{aligned}\mathbb{P}_0(S_{0,i} > t) &= \mathbb{P}_0(S_{0,i} > qi)\\ &\simeq e^{-i\Lambda^*(q)} = e^{-i(\lambda_q q - \Lambda(\lambda_q))}\\ &= e^{-t\left(\lambda_q - \frac{\Lambda(\lambda_q)}{\Lambda'(\lambda_q)}\right)}.\end{aligned}$$

Again

$$\inf_{\lambda:\, \Lambda'(\lambda)>0} \left\{ \lambda - \frac{\Lambda(\lambda)}{\Lambda'(\lambda)} \right\} = \theta^*,$$

where θ^* is the unique positive solution of

$$\Lambda(\lambda) = 0. \tag{2.11}$$

All in all we have

$$\mathbb{P}_0(T > t) \simeq \max_{i \geq 1} \mathbb{P}_0(S_{0,i} > t) \simeq \mathbb{P}_0(S_{0,i^*} > t) \simeq e^{-t\theta^*}, \tag{2.12}$$

where $i^* = t/\Lambda'(\theta^*)$.

2.1.3 Rate functions and concentration inequalities

As we have just worked out, the *upper rate function* r of the process $(S_{0,n})_{n\geq 0}$, i.e.

$$r(q) := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}(S_{0,n} > nq),$$

can be characterized as the *Fenchel-Legendre transform* of the limiting logarithmic moment generating function we defined in (2.10). More precisely, this means that if we define the convex function $\bar{\Lambda}$ as

$$\bar{\Lambda}(\lambda) := \begin{cases} \Lambda(\lambda), & \lambda \geq 0 \\ \infty, & \lambda < 0 \end{cases}$$

then r and $\bar{\Lambda}$ form a pair of two convex conjugate functions on the whole of $\mathbb{R}$. See Figure 2.2 for a picture. The rigorous version of this connection is Hammersley's theorem, which we state in this thesis as Theorem 1 on page 28.

What we needed in our argument was the existence of a unique positive solution θ^* of equation $\Lambda(\lambda) = 0$ which is equivalent to requiring that $\sigma_l := \Lambda'(0+) < 0$. It follows trivially from convex conjugacy that this in turn is equivalent to the statement that $r(q) > 0$ for all $q > \sigma_l$. Putting this together, it follows that θ^* is well defined, if we can show that there exists a value $\sigma_l < 0$ such that for all $q > \sigma_l$ we have $r(q) > 0$.

We will show that in our model $r(q) > 0$ for all $q > \gamma$ (recall that we still assume $\gamma < 0$). One way to prove this is by taking a *concentration inequality*. A concentration inequality gives a statement about how much a distribution is concentrated around its center (usually represented by its mean or its median). The first proof of a concentration inequality for the optimal global score has again been given by Arratia and Waterman in [AW94]. It is based on the Azuma-Hoeffding inequality ([Hoe63] and [Azu67]) and states that for some constant c

$$\mathbb{P}(|S_{0,n} - \mathbb{E}[S_{0,n}]| \geq t) \leq 2 \exp\left(-\frac{t^2}{8c^2 n} \right). \tag{2.13}$$

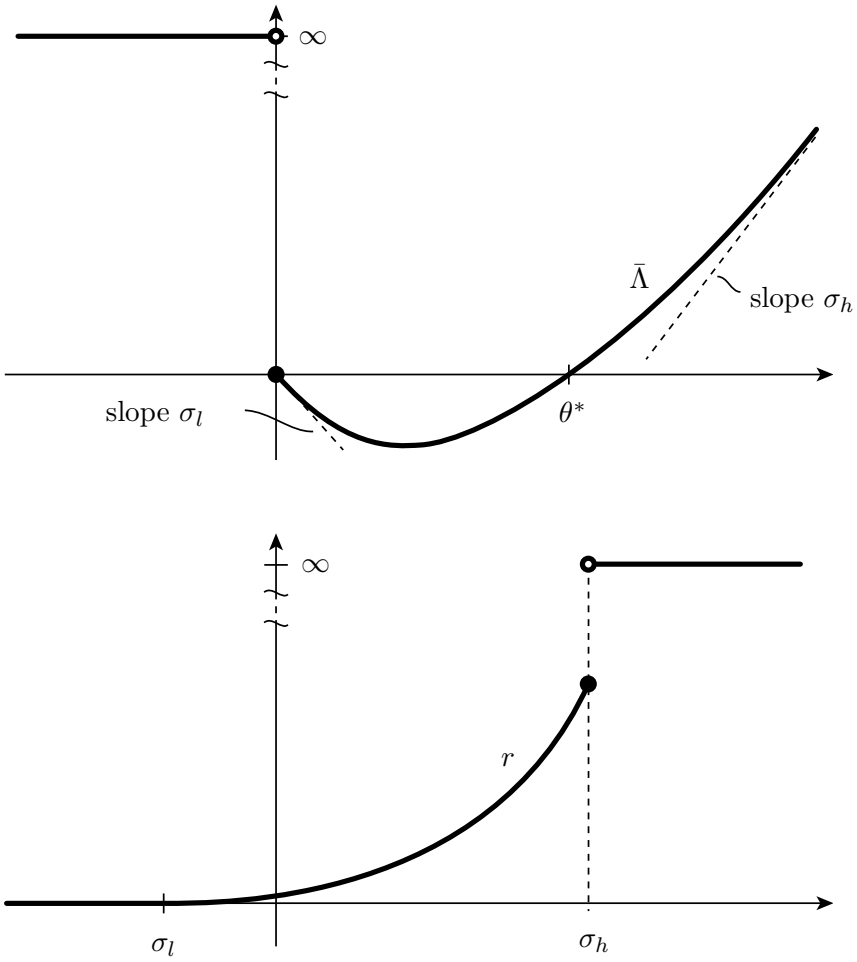

Figure 2.2: *The two convex conjugate functions r and $\bar{\Lambda}$.*

We give the details in Subsection 4.3.1.

In the past years a different technique for proving concentration inequalities has been established by M. Talagrand. It is based on isoperimetric inequalities, and has proved to be very useful in some situations. Especially in the longest increasing subsequence problem, the existing concentration inequalities could be strongly improved. For more information cf. Talagrand's nice article [Tal96] or the relevant parts of Steele's book [Ste97].

The Talagrand technique is also easily applicable to our model. Unfortunately, it does not give a better result than the Azuma-Hoeffding technique. For completeness we give the result in Subsection 4.3.2.

It is immediate that the concentration inequality (2.13) is strong enough to show that $r(q) > 0$ for all $q > \gamma$. It basically says that the fluctuations of $S_{0,n}$ around its expectation are at most of the order of $n^{1/2}$. We give the

exact proof in Section 4.4.

Observe that in general it does not follow from Hammersley's theorem that $r(q) > 0$ for all $q > \gamma$. In fact, the negative of the volume of the Wiener sausage in three or higher dimensions forms an example of an independent superadditive process for which the statement does not hold. For more details on this example and further references confer the survey article of Grimmett ([Gri84]).

2.2 Convergence rates

The fundament for the results we have described up to now came from classical superadditivity theory. Optimal *global* scores can be interpreted as an independent superadditive process, on which we applied the superadditive ergodic theorem and Hammersley's large deviations result. A different family of general techniques came in with the concentration inequalities, which helped us to see that the situation in Hammersley's theorem is especially nice for our model. We then only had to use some measure transformations together with some convexity arguments to get our new large deviations result for the optimal *local* score.

Now more model-specific techniques come into play. General superadditivity arguments provided us with two limiting objects, namely the growth constant

$$\gamma = \lim_{n\to\infty} \frac{\mathbb{E}[S_{0,n}]}{n}$$

and the limiting logarithmic moment generating function

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(\lambda). \tag{2.14}$$

We will now give two results about the rate of convergence in these limiting procedures. A task that has to use more model-specific techniques, since, to quote M. Steele ([Ste97], p. 13),

> *"Until recently, conventional wisdom held that in a problem where a limit is obtained by subadditivity arguments, one has little hope of saying anything about the rate of convergence."*

The *'until recently'* here refers to results which show that for concrete models it sometimes is possible to get results of the rate of convergence, if the model has in fact some additivity-related structure

This has been first done by K. S. Alexander in [Ale93] for the time constant in first-passage percolation and then also for the long-common-subsequence problem in [Ale94]. His results can also be easily extended to our model where we get that a constant c exists such that

$$\mathbb{E}[S_{0,n}] \le n\gamma \le \mathbb{E}[S_{0,n}] + c(n\log n)^{1/2}.$$

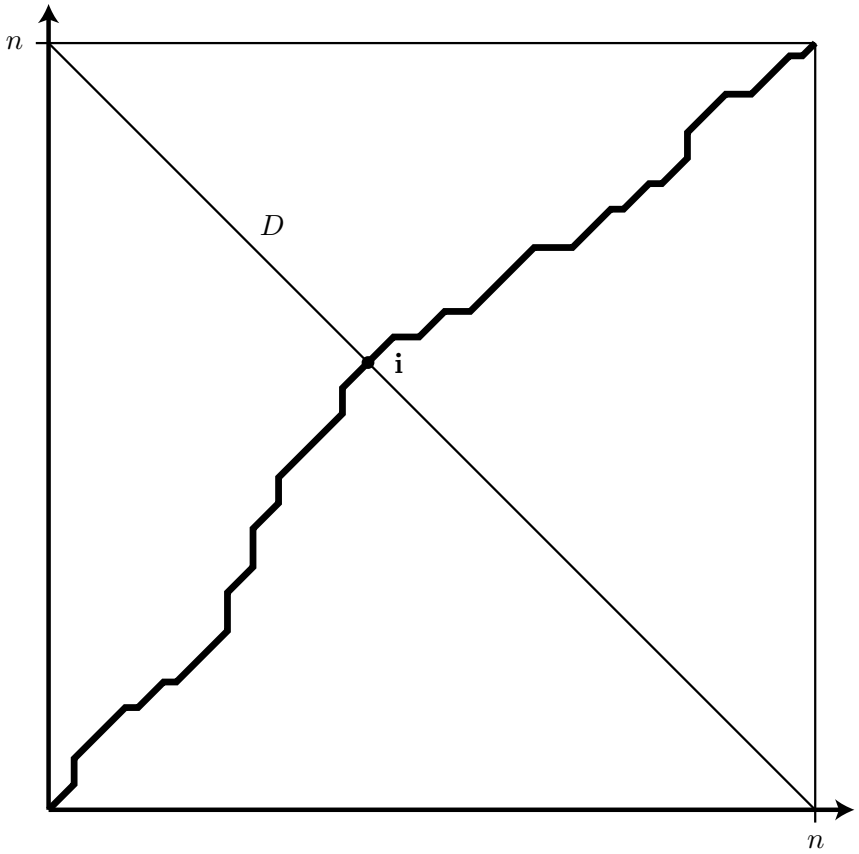

Figure 2.3: *Each optimal path from the origin to* (n, n) *has to cross the diagonal* D *somewhere. This can be used to get some kind of subadditivity structure.*

The proof can be found in Section 4.8 and is essentially an adaptation of the corresponding proof of the long-common-subsequence problem.

For the rate of convergence in (2.14) we have a new result which basically says that

$$\Lambda(\lambda) - \frac{1}{n}\Lambda_n(\lambda) = O\left(\frac{\log n}{n}\right). \tag{2.15}$$

In fact, we have two different proofs of this result, which is stated as Theorem 12 in Section 4.9 and as Corollary 5 in Section 4.10. The first proof is slightly easier than the second, but it also gives slightly worse bounds.

As already mentioned, all the convergence rate results here use the fact that, apart from its superadditivity, the optimal global score also has some additivity-related structure. To explain it cf. Figure 2.3. Take the optimal path for $S_{0,n}$ and the secondary diagonal D. It is clear that the path has to cross D at some vertex $\mathbf{i}$. Denote by $S^{\mathbf{i}}_{0,n}$ the optimal score of all paths from the origin to (n, n) which go through the vertex $\mathbf{i}$. Then $S^{\mathbf{i}}_{0,n}$ can be

split into two independent parts

$$S^{\mathbf{i}}_{0,n} = S_{(0,0),\mathbf{i}} + S_{\mathbf{i},(n,n)},$$

where we use an obvious notation. Maximizing over all vertices on D we get

$$S_{0,n} = \max_{\mathbf{i}\in D}\{S_{(0,0),\mathbf{i}} + S_{\mathbf{i},(n,n)}\}. \tag{2.16}$$

A similar result can of course be formulated for any two points with an arbitrary secondary diagonal between them. Also one can consider more than one diagonal and split the path at each of them. Then one has to extend the maximization in (2.16) to run over all possible combinations of splitting points.

The basic result of splitting at a single diagonal is stated properly in Lemma 4 on page 46. What we concealed up to now for simplicity reasons is the fact that the splitting can not be done exactly the way we have claimed. Due to the special geometry of our lattice there are paths which cross the diagonal without hitting one of the vertices on it. Of course this pitfall is respected in Lemma 4 and all other rigorous applications of the splitting in this thesis.

Chapter 3

The global maximum of an independent superadditive processes with negative drift

Consider a random walk on $\mathbb{R}$ with negative drift, but a positive probability of taking positive steps. It is well-known that, under some regularity conditions, the global maximum of such a process has a distribution with an exponentially decaying tail. The rate of this exponential decay is given by the zero of the logarithmic moment generating function of the step distribution. In this chapter we will show a similar result for independent superadditive processes. The main difference to the random walk case is that the tail's rate of decay will now be characterized as the zero of some *limiting* logarithmic moment generating function.

Before, we treat large deviations for general independent superadditive processes in some breadth, since we do not consider the classical texts to be very satisfying. The primal reference [Ham74] is a good text but written down bottom-up, which makes it difficult to extract the essence of the argument. The next reference is [Kin75], where unfortunately a small error crept in. A more recent, surveying text on the subject is [Gri84]. What motivated us to go that far into details, was the question about the relations of Hammersley's result to the more abstract Gärtner-Ellis-Theorem (cf. [DZ93] or [dH00]). We hope that our exposition below will bring some light into this relation.

3.1 Independent superadditive processes

Let $(T_{i,j})_{0\leq i\leq j}$ be an independent superadditive process, i.e. it fulfils

(i) For all $i \leq j \leq k$

$$T_{i,k} \geq T_{i,j} + T_{j,k},$$

(ii) the joint distribution of $(T_{i,j})_{0\le i\le j}$ is the same as that of $(T_{i+1,j+1})_{0\le i\le j}$,

(iii)
$$\mathbb{E}[T_{0,1}^-] < \infty,$$

(iv) for any increasing sequence $0 = i_0 < i_1 < i_2 < \ldots < i_k$ the random variables $(T_{i_{j-1},i_j})_{1\le j\le k}$ are independent.

Of course, this is more than enough for a superadditive ergodic theorem to hold, i.e. we have (cf. e.g. [Dur96], Section 6.6)

$$\lim_{n\to\infty} \frac{T_{0,n}}{n} = \gamma \quad \text{a.s.}, \tag{3.1}$$

where

$$\gamma := \lim_{n\to\infty} \frac{\mathbb{E}[T_{0,n}]}{n} = \sup_{n>0} \frac{\mathbb{E}[T_{0,n}]}{n} \in (-\infty, \infty]. \tag{3.2}$$

3.2 Large deviations

The important large deviations result for independent superadditive processes is the following

Theorem 1. *For each n define the logarithmic moment generating function Λ_n of $T_{0,n}$ on $\mathbb{R}_+$ as*

$$\Lambda_n(\lambda) := \log \mathbb{E}[\exp(\lambda T_{0,n})].$$

Then the limits

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(\lambda) = \sup_{n>0} \frac{1}{n}\Lambda_n(\lambda) \tag{3.3}$$

and

$$r(q) = \lim_{n\to\infty} \left\{ -\frac{1}{n} \log \mathbb{P}(T_{0,n} > qn) \right\} = \inf_{n>0} \left\{ -\frac{1}{n} \log \mathbb{P}(T_{0,n} > qn) \right\} \tag{3.4}$$

exist in $\bar{\mathbb{R}} := [-\infty, \infty]$ for $\lambda \ge 0$ and $q \in \mathbb{R}$. Moreover, Λ and r are convex functions and are related by

$$r(q) = \sup_{\lambda\ge 0}\{\lambda q - \Lambda(\lambda)\} \tag{3.5}$$

and

$$\Lambda(\lambda) = \sup_{q\in\mathbb{R}}\{q\lambda - r(q)\} =: r^*(\lambda) \tag{3.6}$$

for all q in the interior of $\mathcal{D}_r := \{q' \colon r(q') < \infty\}$ and $\lambda \ge 0$ in the interior of $\mathcal{D}_{r^*} := \{\lambda' \colon r^*(\lambda') < \infty\}$.

Remark 1. The result has originally been stated for superconvolutive families of distribution functions, which, e.g., can arise from independent subadditive processes. For the original version of the statement and the proof see Hammersley ([Ham74]). Also Kingman gives a proof of the result in ([Kin75], Theorem 3.4, p. 214), but his proof, even so he closely follows Hammersley's arguments, is not entirely correct. We will explain this further in Remark 3. In principle, we only translate the proof from [Ham74] to our setting. After doing this, we will discuss the relations to the more modern Gärtner-Ellis-Theorem.

For an illustration how the two functions r and $\bar{\Lambda}$ might look like in a special case, see Figure 2.2 on page 23.

Remark 2. That the two relations (3.5) and (3.6), which connect Λ and r, remind of convex conjugation is not an accident. The only reason we do not characterize Λ and r as a pair of conjugate convex functions is that we do not know whether they are closed (for details about convex conjugacy cf. Appendix A.2). Thus we could have also stated that $\mathrm{cl}(r)$ and $\bar{\Lambda}$ are a pair of convex conjugate functions, where

$$\bar{\Lambda}(\lambda) := \begin{cases} \mathrm{cl}(\Lambda(\lambda)), & \lambda \geq 0 \\ \infty, & \lambda < 0. \end{cases}$$

Proof of Theorem 1. We divide the proof into several steps, to make it more transparent.

Step 1: Let us start with the existence of $\Lambda(\lambda)$ for $\lambda \geq 0$. Observe that, since $(T_{i,j})_{0 \leq i \leq j}$ is an independent superadditive process, we have for $\lambda \geq 0$

$$\begin{aligned} \mathbb{E}[e^{\lambda T_{0,m+n}}] &\geq \mathbb{E}[e^{\lambda(T_{0,m}+T_{m,m+n})}] \\ &= \mathbb{E}[e^{\lambda T_{0,m}}]\mathbb{E}[e^{\lambda T_{m,m+n}}] \\ &= \mathbb{E}[e^{\lambda T_{0,m}}]\mathbb{E}[e^{\lambda T_{0,n}}], \end{aligned}$$

so that after taking logarithms (3.3) follows by superadditivity. Observe especially that for all $n > 0$ and all $\lambda \geq 0$ we have

$$\mathbb{E}[\exp(\lambda T_{0,n})] \leq e^{n\Lambda(\lambda)}.$$

Step 2: The existence of $r(q)$ for general $q \in \mathbb{R}$ follows in much the same way.

$$\begin{aligned} \mathbb{P}(T_{0,m+n} > x + y) &\geq \mathbb{P}(T_{0,m} + T_{m,m+n} > x + y) \\ &\geq \mathbb{P}(T_{0,m} > x, T_{m,m+n} > y) \\ &= \mathbb{P}(T_{0,m} > x)\mathbb{P}(T_{0,n} > y). \end{aligned} \tag{3.7}$$

Replacing x by mq and y by nq in this inequality and taking logarithms gives (3.4) by subadditivity. Of course here we get $r(q) = \infty$ if $\mathbb{P}(T_{0,n} > nq) = 0$ for all n. Also we want to record that

$$\mathbb{P}(T_{0,n} > nq) \leq e^{-nr(q)}. \tag{3.8}$$

Step 3: In this step we show that r is convex. To do this we fix an $\theta \in (0,1)$ and take m and n, which we both let go to infinity such that

$$\frac{m}{m+n} \longrightarrow \theta.$$

Replacing x by mx and y by ny in (3.7) gives after some rearranging

$$-\frac{1}{m+n}\log \mathbb{P}\left(T_{0,m+n} > (m+n)\left[\frac{m}{m+n}x + \frac{n}{m+n}y\right]\right) \leq$$
$$\leq \frac{m}{m+n}\left\{-\frac{1}{m}\log \mathbb{P}(T_{0,m} > mx)\right\} + \frac{n}{m+n}\left\{-\frac{1}{n}\log \mathbb{P}(T_{0,n} > ny)\right\}.$$

Taking the limit we get

$$r(\theta x + (1-\theta)y) \leq \theta r(x) + (1-\theta)r(y),$$

i.e. the convexity of r.

Step 4: We show that $r(q) \geq \sup_{\lambda \geq 0}\{\lambda q - \Lambda(\lambda)\}$. Take any q, any λ and any n and observe that

$$\begin{aligned} e^{n\Lambda(\lambda)} &\geq \mathbb{E}[e^{\lambda T_{0,n}}] \geq \mathbb{E}[e^{\lambda T_{0,n}}; T_{0,n} \geq nq] \\ &\geq e^{\lambda nq}\mathbb{P}(T_{0,n} \geq nq). \end{aligned}$$

Rearranging, taking logarithms and dividing by $-n$ gives

$$-\frac{1}{n}\log \mathbb{P}(T_{0,n} \geq nq) \geq \lambda q - \Lambda(\lambda).$$

Taking the limit $n \to \infty$ yields for all q and $\lambda \geq 0$

$$r(q) \geq \lambda q - \Lambda(\lambda).$$

And finally from a maximization over λ we get that

$$r(q) \geq \sup_{\lambda \geq 0}\{\lambda q - \Lambda(\lambda)\}.$$

From the geometrical interpretation behind convex conjugation it is clear that for all $q \geq \Lambda'(0-)$ we have

$$\Lambda^*(q) := \sup_{\lambda \in \mathbb{R}}\{\lambda q - \Lambda(\lambda)\} = \sup_{\lambda \geq 0}\{\lambda q - \Lambda(\lambda)\} \leq r(q).$$

Again after convex conjugation this gives $r^*(\lambda) \leq \Lambda(\lambda)$ for all $\lambda \geq 0$.

Step 5: In this step we show that $\mathbb{E}[e^{\lambda T_{0,n}}]$ is finite in the interior of $\mathcal{D}_{r^*}$.

We start with some observations on r. From all we know up to now, r is nonnegative, nondecreasing and convex. This already implies that there is a value $\sigma \geq 0$ such that

$$r(q) \sim \sigma q, \quad \text{as } q \to \infty. \tag{3.9}$$

Also we introduced independent superadditive processes in such a way that $\gamma > -\infty$. where γ is the growth constant defined in (3.2). From (3.1) it follows thus that $r(q) = 0$ for all $q < \gamma$.

It is clear from the principle of convex conjugacy, that $r^*(\lambda) = \infty$ for $\lambda > \sigma$ and from Step 4 we also get that $\Lambda(\lambda) = \infty$ for $\lambda > \sigma$. So let us assume that $\sigma > 0$ and take some $0 < \lambda < \sigma$. Also by convex conjugation we have that $r^*(\lambda') < \infty$ for all $0 \leq \lambda' \leq \sigma$ and it follows that λ is in the interior of $\mathcal{D}_{r^*}$.

To argue that also $\mathbb{E}[e^{\lambda T_{0,n}}] < \infty$ for our chosen λ observe first that because of (3.8)

$$0 \leq e^{\lambda x}\mathbb{P}(T_{0,n} > x) \leq e^{\lambda x}e^{-nr(x/n)} = e^{-x\left[\frac{r(x/n)}{x/n} - \lambda\right]}.$$

The crucial thing here is on one hand we know because of (3.9) that

$$\frac{r(x/n)}{x/n} - \lambda \longrightarrow \sigma - \lambda, \qquad x \to +\infty,$$

whereas, because of $r(q) = 0$ for all $q < \gamma$, we get that

$$\frac{r(x/n)}{x/n} - \lambda \longrightarrow -\lambda, \qquad x \to -\infty.$$

Relating the improper integral $\mathbb{E}[e^{\lambda T_{0,n}}]$ via partial integration to the tail of $T_{0,n}$, some easy calculations show that this is precisely what is needed to get the finiteness of $\mathbb{E}[e^{\lambda T_{0,n}}]$.

Step 6: We will now come to the last and crucial step which shows that for $\lambda \geq 0$ in the interior of $\mathcal{D}_{r^*}$ we have $r^*(\lambda) \geq \Lambda(\lambda)$.

As we have discussed in Step 5, the finiteness of $\mathbb{E}[e^{\lambda T_{0,n}}]$ allows to partially integrate it to get

$$\mathbb{E}[e^{\lambda T_{0,n}}] = \int_{-\infty}^{\infty} \lambda e^{\lambda x}\mathbb{P}(T_{0,n} > x)dx = n\lambda \int_{-\infty}^{\infty} e^{n\lambda y}\mathbb{P}(T_{0,n} > ny)dy.$$

The main ingredient is inequality (3.8) which gives that

$$e^{n\lambda y}\mathbb{P}(T_{0,n} > ny) \leq e^{ny[\lambda - r(y)/y]}.$$

This inequality can be exploited in several ways. First observe that for any $y < \gamma$ we have that

$$e^{n\lambda y}\mathbb{P}(T_{0,n} > ny) \leq e^{ny\lambda}. \tag{3.10}$$

Second observe that from (3.9) we also get

$$\lambda - \frac{r(y)}{y} \to \lambda - \sigma < 0, \quad \text{as } y \to \infty.$$

So we can take $0 < \varepsilon < \sigma - \lambda$ and choose a b large enough such that for all $y \geq b$ we have

$$e^{n\lambda y}\mathbb{P}(T_{0,n} > ny) \leq e^{-ny\varepsilon}. \tag{3.11}$$

Third we get from the definition of r^* that

$$e^{n\lambda y}\mathbb{P}(T_{0,n} > ny) \leq e^{nr^*(\lambda)}. \tag{3.12}$$

Putting this together we get

$$\begin{aligned}\mathbb{E}[e^{\lambda T_{0,n}}] &\leq \lambda n\left\{\int_{-\infty}^{a} e^{ny\lambda}dy + \int_a^b e^{nr^*(\lambda)}dy + \int_b^\infty e^{-ny\varepsilon}dy\right\} \\ &= e^{n\lambda a} + n\lambda(b-a)e^{nr^*(\lambda)} + \frac{\lambda}{\varepsilon}e^{-n\varepsilon b}\end{aligned}$$

which after taking logs, dividing by n and letting $n \to \infty$ gives

$$\Lambda(\lambda) \leq \max\{\lambda a, r^*(\lambda), -\varepsilon b\}.$$

Letting $a \to -\infty$ and $b \to \infty$ gives the desired result. □

Remark 3. In [Kin75] Kingman uses essentially the argument we have just presented. The only difference lies in the fact that he only proves (3.6) for $\lambda \geq 0$ from the interior of $\mathcal{D}_\Lambda := \{\lambda' \colon \Lambda(\lambda') < \infty\}$. This set is strictly smaller than the interior of $\mathcal{D}_{r^*}$ in the situation shown in Figure 3.1. Observe that although we have shown in Step 5 that $\Lambda_n(\lambda) = \log \mathbb{E}[\exp(\lambda T_{0,n})]$ is finite for $\lambda \geq 0$ from the interior of $\mathcal{D}_{r^*}$, this still does not imply the finiteness of $\Lambda(\lambda)$. Still, in his formulation of the result Kingman claims that (3.5) holds for all q in $\mathcal{D}_r$. This is obviously not true in this special situation.

Remark 4. It is clear that (iii) above only assures that $\mathbb{E}[T_{0,1}] \in (-\infty, \infty]$. This is also the reason why the finiteness of γ is not clear without further assumptions. Nevertheless, already the requirement that $\Lambda_n(\lambda) < \infty$ for all $0 \leq \lambda \leq l$ in Theorem 1 is enough to assure $\gamma < \infty$.

We will now examine the relations of Theorem 1 to the Gärtner-Ellis-Theorem. For a proof of the Gärtner-Ellis-Theorem cf. [DZ93], Theorem 2.3.6, p. 45 or [dH00], Theorem V.6, p. 54. We will state the result as needed for our purposes.

Let (Z_n) be a sequence of random variables. Define

$$\phi_n(\lambda) := \mathbb{E}[e^{\lambda Z_n}]$$

and assume that

$$\Lambda(\lambda) := \lim_{n\to\infty} \frac{1}{n} \log \phi_n(n\lambda) \text{ exists in } \bar{\mathbb{R}} \tag{GE1}$$

$$0 \in \operatorname{int}(\mathcal{D}_\Lambda), \text{ where } \mathcal{D}_\Lambda := \{\lambda' \in \mathbb{R} \colon \Lambda(\lambda') < \infty\}. \tag{GE2}$$

Then we have

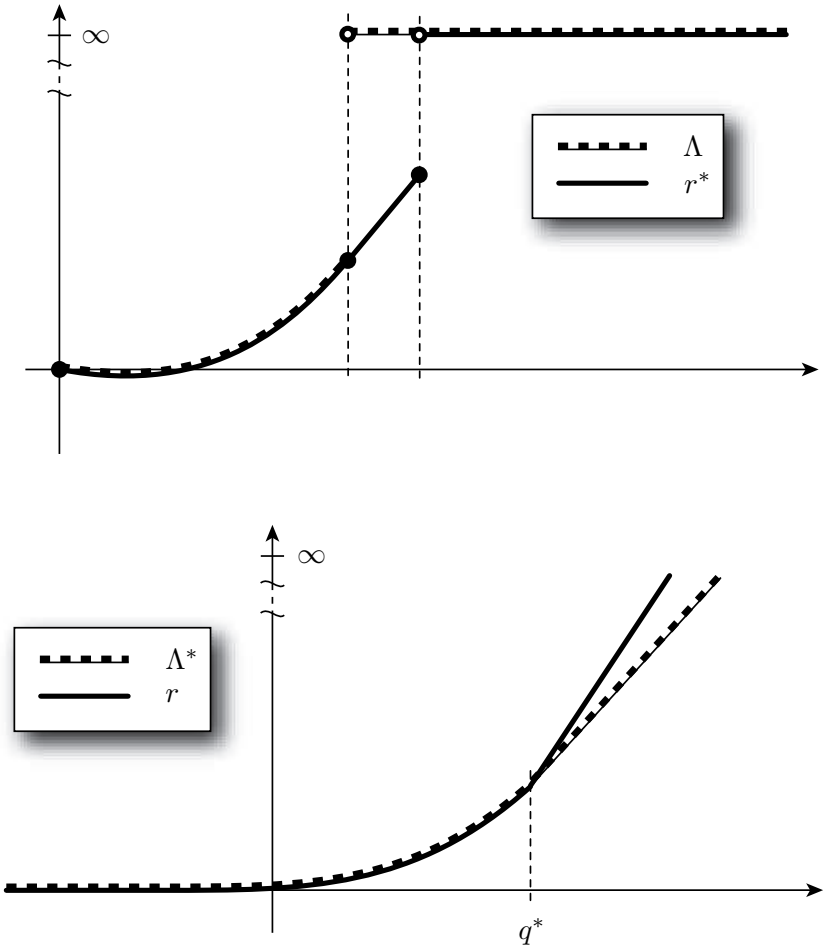

Figure 3.1: *The situation in which Kingman's proof becomes erroneous.*

(i) For all closed $C \subset \mathbb{R}$ we have

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Z_n \in C) \leq - \inf_{q\in C} \Lambda^*(q). \tag{3.13}$$

(ii) For all open $O \subset \mathbb{R}$ we have

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Z_n \in O) \geq - \inf_{q\in O\cap E} \Lambda^*(q), \tag{3.14}$$

where $E = E(\Lambda, \Lambda^*)$ is the set of exposed points of Λ^* whose exposing hyperplane belongs to $\mathrm{int}(\mathcal{D}_\Lambda)$.

An *exposed point* of a convex function $\mathbb{R}$ is a point, where the function is strictly convex. First it is clear that for each value of q for which $\Lambda^*(q) < \infty$

there exists a value λ_q such that

$$\Lambda^*(q) := \sup_{\lambda \in \mathbb{R}} \{\lambda q - \Lambda(\lambda)\} = \lambda_q q - \Lambda(\lambda_q).$$

If Λ^* is strictly convex in q, then this value λ_q is not the maximizer in (3.2) for different values $q' \neq q$ in the sense that for all we have

$$\Lambda^*(q') > \lambda_q q - \Lambda(\lambda_q).$$

In this situation we call λ_q an *exposing hyperplane* for q.

We will see in a minute which role the set E plays in our application of the results to independent superadditive processes. In fact, there is a further statement from the Gärtner-Ellis-Theorem which states that if Λ satisfies

(i) Λ is lower semi-continuous on $\mathbb{R}$,

(ii) Λ is differentiable on $\mathrm{int}(\mathcal{D}_\Lambda)$,

(iii) Either $\mathcal{D}_\Lambda = \mathbb{R}$ or Λ is *steep*, i.e., $\lim_{\lambda \to \pm\infty; \lambda \in \mathcal{D}_\Lambda} \Lambda'(\lambda) = \infty$,

then $O \cap E$ can be replaced by O in (3.14). The diction is that then the family $(\mathbb{P}(Z_n \in \cdot))$ satisfies a *large deviations principle* (LDP) on $\mathbb{R}$ with rate n and rate function Λ^*.

Let us now come back to independent superadditive processes. Clearly we set $Z_n := (1/n)T_{0,n}$, so that with our old definition of Λ_n

$$\phi_n(\lambda) = e^{\Lambda_n(\lambda/n)}.$$

For $\lambda \geq 0$ the convergence in (GE1) has been shown in Step 1 of the proof of Theorem 1 by superadditivity arguments. For $\lambda < 0$ convergence follows by subadditivity with the same argument. Although we did not explicitly require it in Theorem 1 let us also assume that (GE2) holds.

We thus can apply (3.13) and (3.14) to the sets $C = [q, \infty)$ and $O = (q, \infty)$ to get the corresponding result for the rate function of the upper tail. This gives

$$\begin{aligned} \inf_{p \geq q} \Lambda^*(p) &\leq \liminf_{n \to \infty} \left\{ -\frac{1}{n} \log \mathbb{P}(T_{0,n} \geq nq) \right\} \\ &\leq \limsup_{n \to \infty} \left\{ -\frac{1}{n} \log \mathbb{P}(T_{0,n} > nq) \right\} \leq \inf_{\substack{p > q \\ p \in E}} \Lambda^*(p). \end{aligned} \tag{3.15}$$

In some sense the worst situation occurs when Λ is piecewise linear as shown in the top part of Figure 3.2. In this case, all we can deduce from the Gärtner-Ellis-Theorem alone is that the two limits lie somewhere in the shaded region in the lower part of Figure 3.2.

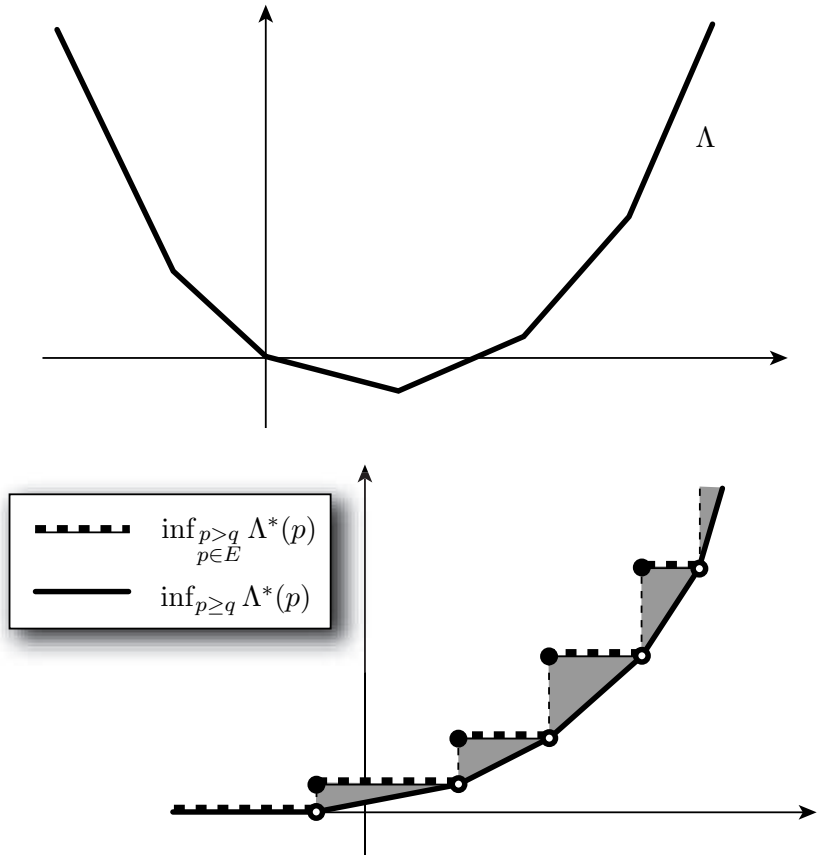

Figure 3.2: *The worst case for the two bounds in* (3.15) *occurs when* Λ *is piecewise linear. From the Gärtner-Ellis-Theorem we only know that the rate function lies somewhere in the shaded region. The rate function, although being nondecreasing, is not necessarily convex.*

The key in the independent superadditive situation lies in the fact that we know much more about the two limits in (3.15). We proved in Step 2 and Step 3 above, that

$$r(q) = \lim_{n\to\infty} \left\{ -\frac{1}{n} \log \mathbb{P}(T_{0,n} > qn) \right\}$$

exists and that r is a convex function. The same can be shown by a similar argument for

$$r'(q) = \lim_{n\to\infty} \left\{ -\frac{1}{n} \log \mathbb{P}(T_{0,n} \geq qn) \right\}.$$

So this extra knowledge, namely the convexity of r and r', immediately implies that

$$r'(q) = r(q) = \Lambda^*(q)$$

as long as there is a point $q' \in E$ with $q' \geq q$. This can be rephrased as $q \in \text{int}(\text{conv}(E))$, where $\text{conv}(E)$ is the *convex hull* of E.

Observe that in general, it is not clear that the limits used to define r and r' exist. If they exist, the functions are clearly nondecreasing, but not necessarily convex. The convexity only followed from the special structure of independent superadditive processes.

Although this looks as if Theorem 1 could be deduced from the Gärtner-Ellis-Theorem together with some convexity arguments, this is not entirely the case. To understand this, consider again the case depicted in the lower part of Figure 3.1. Here Λ^* ends on the right with a linear part. Denote the left beginning of this linear part by q^*. It is clear that then $q^* \notin \text{int}(\text{conv}(E))$, so that we can not deduce the identity of the two functions Λ^* and r for values $q > q^*$. So the same problem that occurs in the Kingman version of Theorem 1 happens to be the weak point, if one wants to prove the result by using the Gärtner-Ellis-Theorem.

3.3 Large deviations for the global maximum

For the global maximum of an independent superadditive process to make sense we need some more requirements. Thus in the following we assume the process $(T_{i,j})_{0 \leq i \leq j}$ to

(i) have a negative growth constant, i.e.

$$\gamma := \lim_{n \to \infty} \frac{\mathbb{E}[T_{0,n}]}{n} = \sup_{n > 0} \frac{\mathbb{E}[T_{0,n}]}{n} < 0,$$

(ii) have a positive probability of being positive, i.e.

$$\mathbb{P}(T_{0,1} > 0) > 0,$$

(iii) be linearly bounded, i.e. there is a constant c such that

$$T_{0,n} \leq cn.$$

In this situation the two functions r and $\bar{\Lambda}$ from Theorem 1 essentially look as shown in Figure 2.2.

Remark 5. Observe that from Theorem 1 it can only be deduced that

$$\gamma \leq \sigma_l := \Lambda'(0+). \tag{3.16}$$

Moreover, it is not even clear that $\sigma_l < 0$.

Remark 6. If it can be shown that $r(q) > 0$ for all $q > \gamma$ then equality in (3.16) follows immediately.

Remark 7. If there is a unique positive solution θ^* of $\Lambda(\lambda) = 0$, then it follows that $\sigma_l < 0$ and also that $r(0) > 0$. This will be useful below.

Remark 8. From the linear boundedness condition it immediately follows that $(1/n)\Lambda_n(\lambda)$ is bounded from above by λc for each n. Together with the convexity of Λ this implies the existence of a limiting slope σ_h of $\Lambda(\lambda)$ for $\lambda \to \infty$.

We define the global maximum of the process $(T_{0,n})_{n\geq 0}$ as

$$G := \max_{n\geq 0} T_{0,n}. \tag{3.17}$$

It is clear by (3.1) that $G < \infty$ a.s. The large deviations of G are given by the following result.

Theorem 2. *Assume that $\Lambda(\lambda) = 0$ has a unique positive solution θ^*. Then*

$$\lim_{t\to\infty} -\frac{1}{t}\log \mathbb{P}(G > t) = \theta^*.$$

Before proving the theorem we need a couple of lemmas.

Lemma 1. *Assume that $\Lambda(\lambda) = 0$ has a unique positive solution θ^*. Then*

$$\theta^* = \inf_{q>0} \frac{r(q)}{q} = \frac{r(q^*)}{q^*}$$

for any value q^* such that

$$\Lambda'(\theta^*-) \leq q^* \leq \Lambda'(\theta^*+).$$

Proof. By convexity and since $r(q) = 0$ for $q \leq \gamma < 0$ it is clear that r is nonnegative and nondecreasing and that we can define

$$\bar{\theta} := \max\{\theta \mid \forall q > 0\colon \theta q \leq r(q)\}$$

which is the slope of the unique increasing straight line that goes through the origin and is tangential to r. On one hand it is clear that

$$\bar{\theta} = \inf_{q>0} \frac{r(q)}{q},$$

since r, as the convex conjugate of $\bar{\Lambda}$, is a closed convex function. On the other hand, since $\bar{\Lambda}$ is the convex conjugate of r, it is clear that $\bar{\Lambda}(\bar{\theta}) = 0$, hence $\bar{\theta} = \theta^*$. □

Lemma 2. *Assume that $\Lambda(\lambda) = 0$ has a unique positive solution θ^*. For all n, all λ such that $\Lambda_n'(\lambda-) > 0$ and all q with*

$$\Lambda_n'(\lambda-) \leq qn \leq \Lambda_n'(\lambda+),$$

we have

$$\theta^* \leq \lambda - \frac{\frac{1}{n}\Lambda_n(\lambda)}{q}.$$

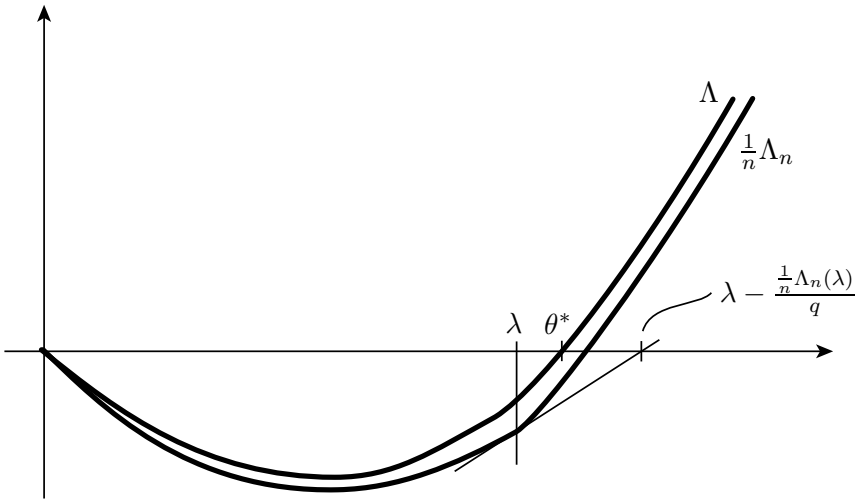

Figure 3.3: *Proof of Lemma 2.*

Proof. Clear from Figure 3.3. □

Lemma 3. *Assume that $\Lambda(\lambda) = 0$ has a unique positive solution θ^*. Then*

$$\lim_{t\to\infty} \min_{n\geq 0} -\frac{1}{t} \log \mathbb{P}(T_{0,n} > t) = \theta^*.$$

Proof. For given t and n define $q_{t,n} = t/n$ and $\lambda_{t,n}$ as the unique value with

$$\frac{1}{n}\Lambda_n'(\lambda_{t,n}-) \leq q_{t,n} \leq \frac{1}{n}\Lambda_n'(\lambda_{t,n}+). \tag{3.18}$$

(Of course, for t and n such that $q_{t,n} > \sigma_h$ this cannot be fulfilled, since

$$\frac{1}{n}\Lambda_n'(\lambda+) \leq \sigma_h,$$

but then $\mathbb{P}(T_{0,n} > t) = 0$ and (3.20) below is trivially fulfilled.)

The basic calculation is

$$\begin{aligned}\mathbb{P}_0(T_{0,n} > t) &= \mathbb{E}_0[1; T_{0,n} > t] \\ &= e^{-t\left[\lambda_{t,n} - \frac{1}{n}\Lambda_n(\lambda_{t,n})/q_{t,n}\right]} \mathbb{E}_{\lambda_{t,n},n}[\exp(-\lambda_{t,n}(T_{0,n} - t)); T_{0,n} > t], \end{aligned} \tag{3.19}$$

where the measure $\mathbb{P}_{\lambda,n}$ is defined via

$$\frac{\mathrm{d}\mathbb{P}_{\lambda,n}}{\mathrm{d}\mathbb{P}}(\omega) := \exp(\lambda T_{0,n}(\omega) - \Lambda_n(\lambda)).$$

Since the expectation in (3.19) is smaller than one it follows for arbitrary n from Lemma 2 that

$$\mathbb{P}_0(T_{0,n} > t) \leq e^{-t\left[\lambda_{t,n} - \frac{1}{n}\Lambda_n(\lambda_{t,n})/q_{t,n}\right]} \leq e^{-t\theta^*}. \tag{3.20}$$

Maximizing over n, taking logarithms and dividing by $-1/t$ gives

$$\lim_{t\to\infty} \min_{n\geq 0} -\frac{1}{t}\log \mathbb{P}(T_{0,n} > t) \geq \theta^*.$$

For the reverse inequality choose q^* such that

$$\Lambda'(\theta^*-) \leq q^* \leq \Lambda'(\theta^*+)$$

and define $n_t^* := \lceil t/q^* \rceil$. Then it is clear that

$$\mathbb{P}(T_{0,n_t^*} > t) \geq \mathbb{P}(T_{0,n_t^*} > n_t^* q^*)$$

or

$$\begin{aligned}\lim_{t\to\infty} \min_{n\geq 0} -\frac{1}{t}\log \mathbb{P}(T_{0,n} > t) &\leq \lim_{t\to\infty} -\frac{1}{t}\log \mathbb{P}(T_{0,n_t^*} > t) \\ &\leq \frac{1}{q^*} \lim_{n_t^*\to\infty} \left\{ \frac{n_t^* q^*}{t} \cdot \left[-\frac{1}{n_t^*}\log \mathbb{P}(T_{0,n_t^*} > n_t^* q^*) \right] \right\} \\ &= \frac{r(q^*)}{q^*} = \theta^*.\end{aligned}$$

□

Observe that from (3.20) we get the following

Corollary 1.

$$\max_{n\geq 0} \mathbb{P}(T_{0,n} > t) \leq e^{-t\theta^*}.$$

It would be interesting to also have a lower bound for the asymptotic behaviour of $\max_{n\geq 0} \mathbb{P}(T_{0,n} > t)$. A first useful step into this direction would be to show that there exists some constant c and some exponent $\eta > 0$ such that

$$\max_{n\geq 0} \mathbb{P}(T_{0,n} > t) \geq \frac{c}{t^\eta} e^{-t\theta^*}.$$

In principle such a result can be deduced from equality (3.19) by controlling the rates of convergence of the terms on the right. We will discuss this issue in more detail later.

Proof of Theorem 2. The proof is based on

$$\max_{n\geq 0} \mathbb{P}(T_{0,n} > t) \leq \mathbb{P}(\max_{n\geq 0} T_{0,n} > t) \leq \sum_{n=0}^{\infty} \mathbb{P}(T_{0,n} > t). \tag{3.21}$$

The first inequality directly gives that

$$\lim_{t\to\infty} -\frac{1}{t}\log \mathbb{P}(\max_{n\geq 0} T_{0,n} > t) \leq \lim_{t\to\infty} \min_{n\geq 0} -\frac{1}{t}\log \mathbb{P}(T_{0,n} > t) = \theta^*,$$

whereas the second gives

$$\lim_{t\to\infty} -\frac{1}{t}\log \mathbb{P}(\max_{n\geq 0} T_{0,n} > t) \geq \lim_{t\to\infty} -\frac{1}{t}\log \sum_{n=0}^{\infty} \mathbb{P}(T_{0,n} > t).$$

So all that remains to show is that

$$\lim_{t\to\infty} -\frac{1}{t}\log \sum_{n=0}^{\infty} \mathbb{P}(T_{0,n} > t) \geq \theta^*.$$

Choose the value q^* such that

$$\Lambda'(\theta^*-) \leq q^* \leq \Lambda'(\theta^*+)$$

Observe that since we assume that θ^* is the unique positive solution of $\Lambda(\lambda) = 0$ there must be value $\bar{\theta} < \theta^*$ with $\Lambda(\bar{\theta}) < 0$ and a value $\bar{q} < q^*$ with

$$\Lambda'(\bar{\theta}-) \leq \bar{q} \leq \Lambda'(\bar{\theta}+),$$

otherwise we would have $\Lambda(\lambda) = 0$ for all $\lambda \in [0, \theta^*]$. Using $\bar{q}$ we define $\bar{n}_t := \lfloor t/\bar{q} \rfloor$.

We now split the sum in (3.21) into two parts

$$\sum_{n=0}^{\infty} \mathbb{P}(T_{0,n} > t) = \sum_{n=0}^{\bar{n}_t} \mathbb{P}(T_{0,n} > t) + \sum_{n=\bar{n}_t+1}^{\infty} \mathbb{P}(T_{0,n} > t). \tag{3.22}$$

For the first sum on the right we certainly have

$$\sum_{n=0}^{\bar{n}_t} \mathbb{P}(T_{0,n} > t) \leq (\bar{n}_t + 1) \max_{n\geq 0} \mathbb{P}(T_{0,n} > t),$$

so taking logarithms and dividing by $-t$ gives

$$-\frac{1}{t}\log \sum_{n=0}^{\bar{n}_t} \mathbb{P}(T_{0,n} > t) \geq$$
$$\geq -\frac{\log(t/\bar{q} + 2) + \max_{n\geq 0} \log \mathbb{P}(T_{0,n} > t)}{t} \xrightarrow[t\to\infty]{} \theta^*.$$

Let us now handle the second sum on the right side of (3.22).

Here we define $q_{t,n} := t/n$ and $\lambda_{t,n}$ as in (3.18). Recall the basic equality (3.19). For convenience we set

$$\mu_{t,n} := \lambda_{t,n} - \frac{\frac{1}{n}\Lambda_n(\lambda_{t,n})}{q_{t,n}},$$

so that we can write

$$\mathbb{P}(T_{0,n} > t) \leq e^{-t\mu_{t,n}}. \tag{3.23}$$

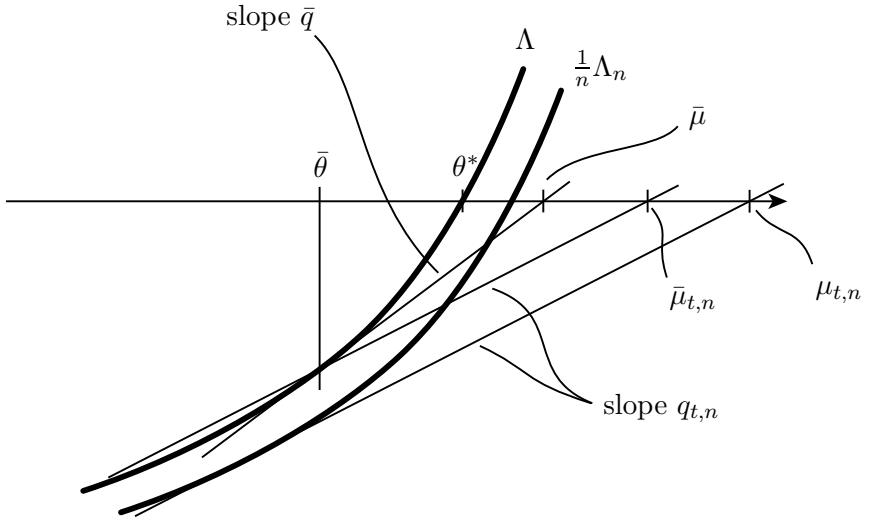

Figure 3.4: *Illustrating* (3.25).

Furthermore it is clear that for the above chosen $\bar{\theta}$ and $\bar{q}$ we have

$$\bar{\mu} := \bar{\theta} - \frac{\Lambda(\bar{\theta})}{\bar{q}} \geq \theta^*.$$

The definition of $\bar{n}_t$ gives that

$$q_{t,\bar{n}_t} = \frac{t}{\bar{n}_t} = \frac{t}{\lfloor t/\bar{q} \rfloor} \geq \bar{q},$$

which implies that

$$\frac{1}{\bar{q}} \geq \frac{\bar{n}_t}{t}. \tag{3.24}$$

Also we have for $n > \bar{n}_t$ that $q_{t,n} \leq \bar{q}$ and it follows from convexity considerations which are clarified in Figure 3.4 that for such n we have

$$\mu_{t,n} = \lambda_{t,n} - \frac{\frac{1}{n}\Lambda_n(\lambda_{t,n})}{q_{t,n}} \geq \bar{\theta} - \frac{\Lambda(\bar{\theta})}{q_{t,n}} =: \bar{\mu}_{t,n}. \tag{3.25}$$

(3.24) and (3.25) together imply that for $n > \bar{n}_t$

$$\mu_{t,n} - \bar{\mu} \geq \frac{\Lambda(\bar{\theta})}{t}(\bar{n}_t - n)$$

or

$$-t(\mu_{t,n} - \bar{\mu}) \leq \Lambda(\bar{\theta})(n - \bar{n}_t). \tag{3.26}$$

finally this gives

$$\begin{aligned}\sum_{n=\bar{n}_t+1}^{\infty} \mathbb{P}(T_{0,n} > t) &\leq e^{-t\bar{\mu}} \sum_{n=\bar{n}_t+1}^{\infty} e^{-t(\mu_{t,n}-\bar{\mu})} \\ &\leq e^{-t\bar{\mu}} \underbrace{\sum_{n=\bar{n}_t+1}^{\infty} e^{\Lambda(\bar{\theta})(n-\bar{n}_t)}}_{=:c_5} \\ &\leq c_5 e^{-t\theta^*}.\end{aligned}$$

□

Again from the proof it is obvious that we have the following upper bound.

Corollary 2. *There exists some constant $c > 0$ such that for t large enough we have*

$$\mathbb{P}(G > t) \leq cte^{-t\theta^*}.$$

Chapter 4

Alignment specific results

This is the main chapter of this thesis. Here, we present everything that we have to say rigorously about optimal alignments of random sequences. We will start in Section 4.1 with a precise definition of global and local optimal alignments and explain some of their properties. In Section 4.3 we will deal with concentration inequalities, i.e. with results about how much the distribution of the optimal global score is spread around its mean.

Next we will come to large deviations to which three sections are devoted. Section 4.4 deals with the large deviations of optimal global scores and is basically an application of the large deviations results of independent superadditive processes presented in Section 3.2. Section 4.5 is the analogue of Section 3.3 and deals with large deviations of the optimal score with a fixed starting point, but a free endpoint. In Section 4.6 the large deviations of optimal local scores are presented.

After having dealt with large deviations, we are able to reprove the phase transition result from [AW94] (together with the completion in [Zha95]). Since we have gained substantially more insight now, we can in Section 4.7 present a better exposition than in the original papers.

Finally we have three sections about the convergence rates of the central limiting objects. These are first the growth constant, where a bound for the convergence rate is given in Section 4.8, and second the limiting logarithmic moment generating function, for which the results can be found in Sections 4.9 and 4.10.

4.1 Optimal alignments and combinatorial optimization

In this section we will define optimal alignments as objects which naturally appear via some combinatorial optimization in a special graphical model. Basically, the model is a version of directed last-passage percolation with some additional features. In last-passage percolation values are assigned to

the edges of some lattice. For each path between two given vertices the sum of the edge-values along the path is calculated. This quantity is then maximized over all paths that connect the two vertices. We will also state those properties of optimal alignments which are directly connected to this graphical representation.

We will start by defining a lattice which we denote by $\mathbb{L}$ and which has vertex set $V_{\mathbb{L}} = \mathbb{N}_0 \times \mathbb{N}_0$. To ease notation we denote points from $\mathbb{N}_0 \times \mathbb{N}_0$ by boldface lower case letters ($\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, ...) and use ordinary lower case letters to refer to their coordinates ($\mathbf{i} = (i_1, i_2)$, $\mathbf{j} = (j_1, j_2)$, ...). Also we use the usual partial ordering on $\mathbb{N}_0 \times \mathbb{N}_0$ that is induced by the ordering on the coordinates, the usual addition and subtraction of points in $\mathbb{Z} \times \mathbb{Z}$ and denote by $||\cdot||_1$ the L_1-norm. We can now define the edge set as

$$E_{\mathbb{L}} = \{e = (\mathbf{i}, \mathbf{j}) \in V_{\mathbb{L}} \times V_{\mathbb{L}} \colon \mathbf{j} - \mathbf{i} \in \{(1,0), (0,1), (1,1)\}\}.$$

We will refer to the different edges of $E_{\mathbb{L}}$ as east-, north- or northeast-edges in an obvious way.

Assign to each edge $e \in E_{\mathbb{L}}$ a value s_e. We start with the northeast-edges, for which the value s_e depends on its location. To do this we consider two fixed sequences $\mathbf{x} = (x_1, x_2, \ldots)$ and $\mathbf{y} = (y_1, y_2, \ldots)$ of letters from some finite alphabet $\mathcal{A}$. To each pair of letters we assign a score using a symmetric function $K \colon \mathcal{A} \times \mathcal{A} \to \mathbb{R}$, called *scoring matrix*. For the northeast-edge $e = (\mathbf{i}, \mathbf{j})$ we set

$$s_e := K(x_{j_1}, y_{j_2})$$

(observe that since e is a northeast-edge we have $j_1 = i_1 + 1$ and $j_2 = i_2 + 1$).

To all the north- and east-edges we simply assign the value of $-\delta$, where $\delta \geq 0$.

An alignment is just a path z through $\mathbb{L}$ that starts at some point $\mathbf{i}$, ends at some point $\mathbf{j} \geq \mathbf{i}$ and is only allowed to take steps to larger vertices. We identify z with the edges it travels through and therefore write $z = (e_i^z)_{1 \leq i \leq |z|}$, where $|z|$ is the length of the alignment.

We now assign to each alignment z a score s_z. The principal ingredient for s_z are the s_{e^z}'s it collects along its way. In addition, we also want to penalize paths for their *roughness*. To do this we subtract the value of $\Delta \geq 0$ for each *horizontal* or *vertical* stretch in the path. This number of horizontal or vertical stretches in the path is obviously a function of the alignment z and we therefore denote it by χ_z. Thus s_z is defined as

$$s_z := \sum_{i=1}^{|z|} s_{e_i^z} - \chi_z \cdot \Delta.$$

We now introduce optimal scores. Denote the set of all alignments which start at $\mathbf{i}$ and end at $\mathbf{j} \geq \mathbf{i}$ by $\mathfrak{Z}_{\mathbf{i},\mathbf{j}}$ and define

$$s_{\mathbf{i},\mathbf{j}} := \max_{z \in \mathfrak{Z}_{\mathbf{i},\mathbf{j}}} s_z.$$

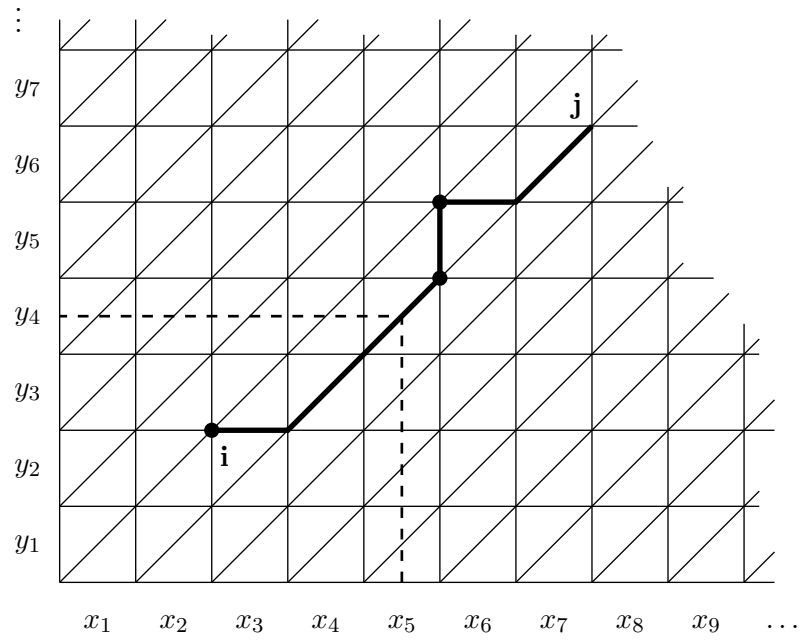

Figure 4.1: *An alignment z from $\mathfrak{Z}_{\mathbf{i},\mathbf{j}}$. The number χ_z of horizontal or vertical stretches in the path is 3. Their starting points are marked by the bullets.*

This is the *optimal global score* of $(x_{i_1+1}, \ldots, x_{j_1})$ and $(y_{i_2+1}, \ldots, y_{j_2})$. As a convention we set $s_{\mathbf{i},\mathbf{i}} = 0$ for all $\mathbf{i}$. The *optimal local score* for the same two parts of the sequences is defined as

$$m_{\mathbf{i},\mathbf{j}} := \max_{\mathbf{i} \le \mathbf{k} \le \mathbf{l} \le \mathbf{j}} s_{\mathbf{k},\mathbf{l}}.$$

We call any alignment $z \in \mathfrak{Z}_{\mathbf{i},\mathbf{j}}$ for which $s_z = s_{\mathbf{i},\mathbf{j}}$ an *optimal global alignment for* $(x_{i_1+1}, \ldots, x_{j_1})$ *and* $(y_{i_2+1}, \ldots, y_{j_2})$ (or simply an *optimal alignment for* $s_{\mathbf{i},\mathbf{j}}$). Analogously we call any alignment $z \in \mathfrak{Z}_{\mathbf{k},\mathbf{l}}$, where $\mathbf{i} \le \mathbf{k} \le \mathbf{l} \le \mathbf{j}$ and for which $s_z = m_{\mathbf{i},\mathbf{j}}$ an *optimal local alignment for* $(x_{i_1+1}, \ldots, x_{j_1})$ *and* $(y_{i_2+1}, \ldots, y_{j_2})$ (or simply an *optimal alignment for* $m_{\mathbf{i},\mathbf{j}}$).

Remark 9. It is clear that all alignments z in $\mathfrak{Z}_{\mathbf{i},\mathbf{j}}$ that hit a certain set of northeast edges have the same number of north- and east-steps between visiting two successive northeast-edges. Of course, the order of these north- and east-steps influences the number of direction changes χ_z. This number is clearly minimal when the north- and east-steps are ordered in such a way that all the north-steps precede the east-steps or vice versa. This implies that when searching for optimal alignments it suffices to consider alignments in which the north- and east-steps are ordered in one of the two ways.

Remark 10. Observe that we use the same letter s with different lower-right indices for different objects, namely for edge-values, scores of alignments and

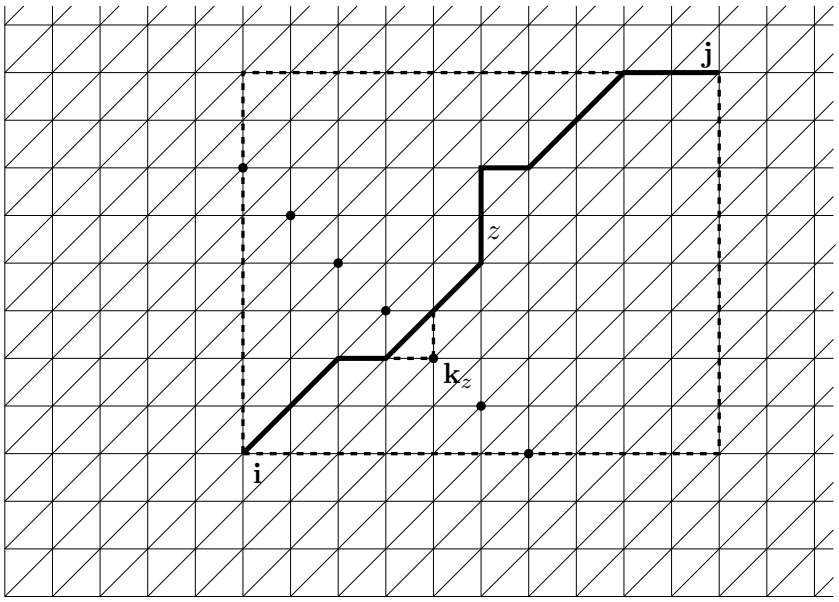

Figure 4.2: *Illustrating Lemma 4. In this case* $||\mathbf{j}-\mathbf{i}||_1 = 18$ *and* $d = 6$.

optimal global scores. Nevertheless, this should not lead to any confusion, since the meaning will always be clear from the index.

The collection $(s_{\mathbf{i},\mathbf{j}})_{(0,0)\leq\mathbf{i}\leq\mathbf{j}}$ has a nice superadditivity structure in the sense that, whenever $\mathbf{i} \leq \mathbf{j} \leq \mathbf{k}$, we have

$$s_{\mathbf{i},\mathbf{k}} \geq s_{\mathbf{i},\mathbf{j}} + s_{\mathbf{j},\mathbf{k}}.$$

In addition, it also has a certain *splitting property* which is described by the following

Lemma 4. *Let* $\mathbf{j} \geq \mathbf{i}$. *Fix some integer* d *with* $0 \leq d \leq ||\mathbf{j}-\mathbf{i}||_1$. *Then there is a constant* c_2 *(not depending on the letter sequences* $\mathbf{x}$ *and* $\mathbf{y}$*) such that*

$$S_{\mathbf{i},\mathbf{j}} - c_2 \leq \max_{\substack{\mathbf{k}\colon \mathbf{i}\leq\mathbf{k}\leq\mathbf{j} \\ ||\mathbf{k}-\mathbf{i}||_1=d}} \{s_{\mathbf{i},\mathbf{k}} + s_{\mathbf{k},\mathbf{j}}\} \leq s_{\mathbf{i},\mathbf{j}}.$$

Proof. In the cases $\mathbf{i} = \mathbf{j}$, $d = 0$ or $d = ||\mathbf{j}-\mathbf{i}||_1$ there is nothing to prove, so let us assume that $\mathbf{j} > \mathbf{i}$ and $0 < d < ||\mathbf{j}-\mathbf{i}||_1$. Also only the first inequality is to prove, since the second directly follows from superadditivity.

For the proof let z be an optimal alignment for $s_{\mathbf{i},\mathbf{j}}$. As shown in Figure 4.2 the vertices $\{\mathbf{k}\colon \mathbf{i} \leq \mathbf{k} \leq \mathbf{j}, ||\mathbf{k}-\mathbf{i}||_1 = d\}$ lie on a diagonal which z has to cross somewhere. Actually z can be modified slightly such that it hits one of the vertices on that diagonal. Denote the modified path by $\tilde{z}$ and the vertex it hits on the diagonal by $\mathbf{k}_z$. It is clear that, even in the worst situation, this modification does not *cost* too much in the sense that

$$s_{\mathbf{i},\mathbf{j}} = s_z \leq s_{\tilde{z}} + [k^* + 2(\Delta + \delta)],$$

where

$$k^* := \max_{\alpha,\beta\in\mathcal{A}} K^+(\alpha,\beta). \tag{4.1}$$

We can split $\tilde{z}$ into two parts by cutting it at $\mathbf{k}_z$. We denote these parts by $\tilde{z}_1$ and $\tilde{z}_2$ resp., where $\tilde{z}_1 \in \mathfrak{Z}_{\mathbf{i},\mathbf{k}_z}$ and $\tilde{z}_2 \in \mathfrak{Z}_{\mathbf{k}_z,\mathbf{j}}$. Again it is clear that in the worst case such a splitting costs

$$s_{\tilde{z}} \le s_{\tilde{z}_1} + s_{\tilde{z}_2} + \Delta.$$

All in all this gives with $c_2 := k^* + 3\Delta + 2\delta$

$$\begin{aligned} s_{\mathbf{i},\mathbf{j}} - c_2 \le s_{\tilde{z}} - \Delta \le s_{\tilde{z}_1} + s_{\tilde{z}_2} \le s_{\mathbf{i},\mathbf{k}_z} + s_{\mathbf{k}_z,\mathbf{j}} \\ \le \max_{\substack{\mathbf{k}\colon \mathbf{i}\le\mathbf{k}\le\mathbf{j} \\ ||\mathbf{k}-\mathbf{i}||_1=d}} \{s_{\mathbf{i},\mathbf{k}} + s_{\mathbf{k},\mathbf{j}}\}. \end{aligned}$$

□

4.2 Introducing randomness

We now introduce randomness into the model by randomizing the underlying letter sequences. On the notational side, the introduction of randomness is reflected by the use of upper case instead of lower case letters, i.e. we write $\mathbf{X}$ instead of $\mathbf{x}$, S_e instead of s_e, $S_{\mathbf{i},\mathbf{j}}$ instead of $s_{\mathbf{i},\mathbf{j}}$, ...

We randomize the letter sequences by drawing all the letters independently according to some distribution μ on the finite alphabet $\mathcal{A}$. Thus the process $(S_{\mathbf{i},\mathbf{j}})_{(0,0)\le\mathbf{i}\le\mathbf{j}}$ becomes an independence superadditive process in the sense of Section 3.1. Observe that now we deal with a two-dimensional index set, so the conditions (i) to (iv) from page 27 have to be translated correspondingly, e.g. (ii) should read as

(ii)' for any unit vector $\mathbf{e} \in \mathbb{N}_0 \times \mathbb{N}_0$ (i.e. $\mathbf{e} = (1,0)$ or $\mathbf{e} = (0,1)$) the joint distribution of $(S_{\mathbf{i},\mathbf{j}})_{(0,0)\le\mathbf{i}\le\mathbf{j}}$ is the same as that of $(S_{\mathbf{i}+\mathbf{e},\mathbf{j}+\mathbf{e}})_{(0,0)\le\mathbf{i}\le\mathbf{j}}$.

Condition (iii) is fulfilled since the process satisfies

$$-2\Delta - ||\mathbf{j}-\mathbf{i}||_1\delta \le S_{\mathbf{i},\mathbf{j}} \le k^* \min(j_1 - i_1, j_2 - i_2) \le k^*||\mathbf{j}-\mathbf{i}||_1/2 \tag{4.2}$$

(recall the definition of k^* in (4.1)).

In considerations related to the growth constant γ of the superadditive process we have to be more careful because of the two-dimensionality of the index set. Whereas in the one-dimensional situation there is only one way to let i go to infinity, there are many directions into which the two-dimensional index $\mathbf{i}$ can go to infinity. To be more specific, let $\rho = \rho_1/\rho_2 \in [-1,1] \cap \mathbb{Q}$, where we require that $\rho_1 \in \mathbb{Z}$ and $\rho_2 \in \mathbb{N}$ and $\gcd(\rho_1,\rho_2) = 1$ (we make the convention $0 = 0/1$). Denote by

$$I_\rho := \{k(\rho_2 + \rho_1, \rho_2 - \rho_1)\colon k \in \mathbb{N}_0\}$$

the set of vertices of $\mathbb{L}$ which lie on the line through the origin with slope $(\rho_2 + \rho_1)/(\rho_2 - \rho_1)$. In analogy to (3.2) we define

$$\gamma_\rho := \lim_{\substack{\mathbf{n}\in I_\rho \\ ||\mathbf{n}||_1\to\infty}} \frac{\mathbb{E}[S_{(0,0),\mathbf{n}}]}{||\mathbf{n}||_1/2} = \sup_{\mathbf{n}\in I_\rho} \frac{\mathbb{E}[S_{(0,0),\mathbf{n}}]}{||\mathbf{n}||_1/2}$$

and have as in (3.1)

$$\lim_{\substack{\mathbf{n}\in I_\rho \\ ||\mathbf{n}||_1\to\infty}} \frac{S_{(0,0),\mathbf{n}}}{||\mathbf{n}||_1/2} = \gamma_\rho \quad \text{a.s.} \tag{4.3}$$

Due to the symmetry of the scoring matrix K and since we use the L_1-norm in the definition of γ_ρ we have

Lemma 5. *For all* $\rho \in [-1, 1] \cap \mathbb{Q}$

$$\gamma_\rho \leq \gamma_0.$$

Proof. Let $\rho \in [-1, 1] \cap \mathbb{Q}$ and $\varepsilon > 0$. Choose $\mathbf{n} \in I_\rho$ large enough such that

$$\frac{\mathbb{E}[S_{(0,0),\mathbf{n}}]}{||\mathbf{n}||_1/2} \geq \gamma_\rho - \varepsilon.$$

Denote by $\bar{\mathbf{n}}$ the vertex one gets by interchanging the coordinates of $\mathbf{n}$. It is clear that $\mathbf{n} + \bar{\mathbf{n}} \in I_0$. Also $||\mathbf{n} + \bar{\mathbf{n}}||_1 = 2||\mathbf{n}||_1$ and $\mathbb{E}[S_{(0,0),\mathbf{n}}] = \mathbb{E}[S_{(0,0),\bar{\mathbf{n}}}]$ by symmetry of K, so we get

$$\gamma_0 \geq \frac{\mathbb{E}[S_{(0,0),\mathbf{n}+\bar{\mathbf{n}}}]}{||\mathbf{n} + \bar{\mathbf{n}}||_1/2} \geq \frac{\mathbb{E}[S_{(0,0),\mathbf{n}} + S_{\mathbf{n},\mathbf{n}+\bar{\mathbf{n}}}]}{||\mathbf{n}||_1} = \frac{\mathbb{E}[S_{(0,0),\mathbf{n}}]}{||\mathbf{n}||_1/2} \geq \gamma_\rho - \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the proof is complete. □

The Lemma shows that the northeast direction in some sense is the dominant direction. We therefore set

$$\gamma := \gamma_0. \tag{4.4}$$

We will later see that for our purposes it suffices to consider the one-dimensional superadditive process we get when we restrict ourselves to the $S_{\mathbf{i},\mathbf{j}}$, where $\mathbf{i}, \mathbf{j} \in I_0$. For convenience we therefore introduce the shorthand notation

$$S_{i,j} := S_{(i,i),(j,j)},$$

where $i, j \in \mathbb{N}_0$ and $0 \leq i \leq j$. $(S_{i,j})_{0\leq i\leq j}$ is then an ordinary one-dimensional independent superadditive process in the sense of Section 3.1.

Also we introduce the same shorthand notation for the optimal local score and write

$$M_{i,j} := M_{(i,i),(j,j)}.$$

4.3 Concentration inequalities

In this section we derive results about the concentration of the distribution of $S_{0,n}$ around its mean. The main result states that the tails of the distribution of $S_{0,n}$ are sub-Gaussian. We will prove this result using two different techniques. One is a martingale technique that goes back to Hoeffding ([Hoe63]) and Azuma ([Azu67]). It has first been applied to optimal global scores by Arratia and Waterman ([AW94]), although they did not formulate the full result for both tails. The second is a concentration inequality from Talagrand ([Tal96]), which gives slightly worse bounds but is an interesting technique for problems from the field of combinatorial optimization and therefore worth mentioning. Its application to optimal global scores is new, although it has been applied to very similar problems as, e.g., the distribution of the longest increasing subsequence of a random sequence (cf. [Tal96]).

4.3.1 Martingale technique

We first state the Azuma-Hoeffding inequality. For a good exposition of the result see the very beginning of Steele's book ([Ste97]).

Definition 1. *A sequence $(D_i)_{1\leq i\leq n}$ of random variables satisfies the* product identity*, iff for any sequences of constants (a_i) and (b_i) we have*

$$\mathbb{E}\left[\prod_{i=1}^{n}(a_i + b_i D_i)\right] = \prod_{i=1}^{n} a_i.$$

Theorem 3. *For a sequence $(D_i)_{1\leq i\leq n}$ of random variables satisfying the product identity we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} D_i\right| \geq t\right) \leq 2\exp\left(-t^2/\left(2\sum_{i=1}^{n}||D_i||_\infty^2\right)\right).$$

Proof. See [Ste97], p. 4. □

The theorem is mainly applied in the situation where the D_i's are martingale differences and the situation is such that the boundedness is clear. Martingale differences obey some kind of orthogonality relation in the sense that for all $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ we have

$$\mathbb{E}\left[\prod_{j=1}^{k} D_{i_j}\right] = 0.$$

This entails the product identity for the D_i's.

Let us now come to the concrete application of Theorem 3 to optimal global scores. To introduce a martingale structure we pair the letters from $\mathbf{X}$ and $\mathbf{Y}$ to a sequence $\mathbf{Z}$ via $Z_i := (X_i, Y_i)$. We then define $\mathcal{F}_k$ to be the sigma-field that is generated by $Z_1, \ldots, Z_k$ for $1 \leq k$, i.e.

$$\mathcal{F}_k := \sigma(Z_1, \ldots, Z_k).$$

We define the D_i's as the differences of the martingale $(\mathbb{E}[S_{0,n}|F_i])_{1\leq i\leq n}$, or

$$D_i := \mathbb{E}[S_{0,n}|F_i] - \mathbb{E}[S_{0,n}|F_{i-1}].$$

It is clear that

$$S_{0,n} - \mathbb{E}[S_{0,n}] = \sum_{i=1}^{n} D_i$$

and the D_i's are bounded by the following

Lemma 6. *There is a constant c_3 such that*

$$||D_i||_\infty \leq 2c_3.$$

Proof. We write $S_{0,n} =: s(Z_1, \ldots, Z_n)$ to express the functional dependence of $S_{0,n}$ from the random letter pairs. Let $\hat{Z}_i$ be a random letter pair that is chosen independently from $(Z_1, \ldots, Z_n)$ but with the same distribution as Z_i. Because of the independence

$$\mathbb{E}[s(Z_1, \ldots, Z_i, \ldots, Z_n)|\mathcal{F}_{i-1}] = \mathbb{E}[s(Z_1, \ldots, \hat{Z}_i, \ldots, Z_n)|\mathcal{F}_i]$$

from which follows that we can write

$$D_i = \mathbb{E}[s(Z_1, \ldots, Z_i, \ldots, Z_n) - s(Z_1, \ldots, \hat{Z}_i, \ldots, Z_n)|\mathcal{F}_i].$$

Let now $z_1, \ldots, z_n$ and $\hat{z}_i$ be any letter pairs from $\mathcal{A} \times \mathcal{A}$. We argue that

$$|s(z_1, \ldots, z_i, \ldots, z_n) - s(z_1, \ldots, \hat{z}_i, \ldots, z_n)| \leq 2c_3, \tag{4.5}$$

where

$$c_3 := \max_{\alpha,\beta\in\mathcal{A}} K(\alpha, \beta) - \min_{\alpha',\beta'\in\mathcal{A}} K(\alpha', \beta')$$

is the maximal difference of entries one can find in the scoring matrix K. Indeed, the worst case happens if in the optimizing alignment for

$$s := s(z_1, \ldots, z_i, \ldots, z_n)$$

both letters from z_i are aligned to different letters. In that case, changing the letter pair z_i to $\hat{z}_i$ can maximally decrease s by $2c_3$. If the letter pairs from z_i are aligned to each other or only one of them is aligned to another letter, s is maximally decreased by c_3, whereas if they are not aligned at all it cannot decrease. Reversing the role of z_i and $\hat{z}_i$ in this argument proves (4.5).

Since (4.5) is true for every realization of $Z_1, \ldots, Z_n$ and $\hat{Z}_i$, the proof is complete. □

We thus have proven the following

Theorem 4. *Whenever the letter pairs Z_i are chosen independently we have*

$$\mathbb{P}(|S_{0,n} - \mathbb{E}[S_{0,n}]| \geq t) \leq 2\exp\left(-\frac{t^2}{8c_3^2 n}\right),$$

where c_3 is the constant from Lemma 6.

4.3.2 Talagrand's concentration inequality

Before applying Talagrand's concentration inequality to optimal global scores, we have to spend some time on a careful formulation of the result.

The abstract setting is given by the n-fold product Ω^n of some probability space Ω. For sets $A \subseteq \Omega^n$ and points $x \in \Omega^n$ we now develop a special notion of *distance*.

To start with we introduce for any two points $x, y \in \Omega^n$ the *Hamming-difference* $h(x,y) \in \mathbb{R}^n$ by

$$h(x,y)_i := \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i. \end{cases}$$

The usual *Hamming-distance* $d_H(x,y)$ could now be defined by the $||\cdot||_1$-norm of $h(x,y)$, i.e.

$$d_H(x,y) := ||h(x,y)||_1$$

and then extended to measure the distance between x and A via

$$d_H(x,A) := \inf_{y \in A} d_H(x,y).$$

However we will define a different notion of distance by first introducing the convex hull $V_A(x)$ of the set of Hamming-differences between x and the points $y \in A$, i.e.

$$V_A(x) := \mathrm{conv}\{h(x,y) \colon y \in A\} \subseteq \mathbb{R}^n$$

and then define

$$d_T(x,A) := \inf_{z \in V_A(x)} ||z||_2.$$

This notion of distance is sometimes referred to as *Talagrand's convex distance* (cf. [Ste97]). The key now lies in the following theorem from [Tal96].

Theorem 5. *Let P be a product measure on Ω^n. Then*

$$\int_{\Omega^n} \exp\left(\frac{1}{4} d_T^2(x,A)\right) dP(x) \leq \frac{1}{P(A)}.$$

In particular

$$P(d_T(\cdot, A) \geq t) \leq \frac{1}{P(A)} \exp\left(\frac{-t^2}{4}\right).$$

Proof. See [Tal96], Theorem 6.1. □

To apply Theorem 5 we need the following Lemma.

Lemma 7. *For any $x \in \Omega^n$ and any $A \subseteq \Omega^n$ we can find $y \in A$ such that*

$$d_H(x, y) \leq \sqrt{n} d_T(x, A).$$

Proof. This is just a special case of Lemma 6.3 in [Tal96]. □

We now can state and prove the analogue of Theorem 4. The appearance of the median instead of the mean is just more convenient in the formulation of the result, but does not produce any complications in practice.

Theorem 6. *Whenever the letter pairs Z_i are chosen independently we have*

$$\mathbb{P}(|S_{0,n} - \operatorname{med}(S_{0,n})| \geq t) \leq 4 \exp\left(-\frac{t^2}{16 c_3^2 n}\right),$$

where c_3 is the constant from Lemma 6.

Proof. In the language of Theorem 5 we set $\Omega = \mathcal{A} \times \mathcal{A}$ and for $x \in \Omega^n$ we denote by $s(x)$ the optimal global alignment score. According to Lemma 7 and with an argument similar to the one in Lemma 6 we have that for all $x \in \Omega^n$ and all $A \subseteq \Omega^n$ there exists a $y \in A$ with

$$s(x) - s(y) \leq 2c_3 \cdot d_H(x, y) \leq 2c_3 \sqrt{n} d_T(x, A). \tag{4.6}$$

In the following define

$$A(a) := \{x \in \Omega^n \colon s(x) \leq a\}.$$

It follows from (4.6) for all $x \in \Omega^n$ with the appropriately chosen $y \in A(a)$ that

$$a \geq s(y) \geq s(x) - 2c_3 \sqrt{n} d_T(x, A(a)).$$

In particular we have

$$s(x) \geq a + t \quad \Longrightarrow \quad d_T(x, A(a)) \geq \frac{t}{2c_3\sqrt{n}}. \tag{4.7}$$

(i) Applying (4.7) to $a = \operatorname{med}(S_{0,n})$ gives

$$s(x) \geq \operatorname{med}(S_{0,n}) + t \quad \Longrightarrow \quad d_T(x, A(\operatorname{med}(S_{0,n}))) \geq \frac{t}{2c_3\sqrt{n}}$$

and we get from Theorem 5 that

$$\begin{aligned}
\mathbb{P}(S_{0,n} - \operatorname{med}(S_{0,n}) \geq t) &\leq \mathbb{P}\left(d_T(\cdot, A(\operatorname{med}(S_{0,n}))) \geq \frac{t}{2c_3\sqrt{n}}\right) \\
&\leq \mathbb{P}(A(\operatorname{med}(S_{0,n})))^{-1} \exp\left\{\frac{-t^2}{16c_3^2 n}\right\} \\
&\leq 2 \exp\left\{\frac{-t^2}{16c_3^2 n}\right\},
\end{aligned}$$

since $\mathbb{P}(A(\operatorname{med}(S_{0,n}))) \geq 1/2$.

(ii) Again applying (4.7), but this time to $a = \mathrm{med}(S_{0,n}) - t$ gives

$$s(x) \geq \mathrm{med}(S_{0,n}) \quad \Longrightarrow \quad d_T(x, A(\mathrm{med}(S_{0,n}) - t)) \geq \frac{t}{2c_3\sqrt{n}}$$

which implies

$$\begin{aligned}\frac{1}{2} &\leq \mathbb{P}(S_{0,n} \geq \mathrm{med}(S_{0,n})) \\ &\leq \mathbb{P}\left(d_T(\cdot, A(\mathrm{med}(S_{0,n}) - t)) \geq \frac{t}{2c_3\sqrt{n}}\right) \\ &\leq \mathbb{P}(A(\mathrm{med}(S_{0,n}) - t))^{-1} \exp\left\{\frac{-t^2}{16c_3^2 n}\right\}\end{aligned}$$

This gives

$$\mathbb{P}(S_{0,n} - \mathrm{med}(S_{0,n}) \leq -t) \leq 2\exp\left\{\frac{-t^2}{16c_3^2 n}\right\}.$$

□

4.4 Large deviations for optimal global scores

As already mentioned in Section 4.2, $(S_{i,j})_{0\leq i\leq j}$ forms an independent superadditive process. We thus can apply the results from Section 3.1, especially Theorem 1. What is important here is that, due to the concentration inequalities from Section 4.3, we can prove the positiveness of the large deviations rate function $r(q)$ for all $q > \gamma$.

Lemma 8. *For the independent superadditive process $(S_{i,j})_{0\leq i\leq j}$ we have for all $q > \gamma$*

$$r(q) > 0,$$

where γ is the growth constant introduced in (4.4) and

$$r(q) := \lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}(S_{0,n} > qn) = \inf_{n>0} -\frac{1}{n}\log\mathbb{P}(S_{0,n} > qn).$$

Proof. Let $q > \gamma$ and set $\varepsilon := q - \gamma$. Observe that because of (3.2) we have for n large enough

$$n(\gamma - \varepsilon) \leq \mathbb{E}[S_{0,n}] \leq n\gamma.$$

This gives

$$\begin{aligned}\mathbb{P}(S_{0,n} > nq) &= \mathbb{P}(S_{0,n} > n\gamma + \varepsilon n) \\ &\leq \mathbb{P}(S_{0,n} - \mathbb{E}[S_{0,n}] \geq \varepsilon n/2) \\ &\leq 2\exp\left(-\frac{\varepsilon^2 n}{64c_3^2}\right),\end{aligned}$$

where we used Theorem 4 for the last inequality. Taking logarithms, dividing by $-n$ and letting $n \to \infty$ gives

$$r(q) > \frac{(q-\gamma)^2}{64c_3^2}.$$

□

This nicely completes the picture we developed in Section 3.1. Since by (4.2) $S_{0,n}$ is linearly bounded in n, the function $\Lambda(\lambda)$ from Theorem 1 has a limiting slope σ_h as $\lambda \to \infty$. On the other hand the last Lemma proves that $\Lambda'(0+) = \gamma$. So we see that Λ does not look too pathologically. Especially it follows that in the case $\gamma < 0$ there is a unique positive solution θ^* of the equation $\Lambda(\lambda) = 0$ iff $\sigma_h > 0$.

4.5 Maximizing with a fixed starting point

In this section we will prove the analogue of the general results from Section 3.3 in the world of optimal alignments. To do so, we require in the following that $(S_{i,j})_{0 \le i \le j}$ fulfils the conditions stated at the beginning of Section 3.3. These are

$$\gamma < 0$$

and

$$\mathbb{P}(S_{0,1} > 0) > 0.$$

Recall that the third condition is already fulfilled by (4.2). These conditions are sufficient to assure that the equation $\Lambda(\lambda) = 0$ has a unique positive solution θ^*.

Of course, if we kept on restricting ourselves to the one-dimensional process $(S_{i,j})_{0 \le i \le j}$ we could directly apply the results from Section 3.3. But instead of defining G as in (3.17), we want to give credit to the two-dimensionality of the problem by defining

$$G := \max_{\mathbf{j} \ge (0,0)} S_{(0,0),\mathbf{j}}. \tag{4.8}$$

Anyway, this hardly changes the main result

Theorem 7.

$$\lim_{t \to \infty} -\frac{1}{t} \log \mathbb{P}(G > t) = \theta^*.$$

Proof. The statement can essentially be proved by following the line of argument for Theorem 2. We will only point out how to handle the problems that arise from the two-dimensionality.

To extend notation we first define for all $\mathbf{i} \in \mathbb{L}_V$

$$\Lambda_{\mathbf{i}}(\lambda) := \log \mathbb{E}[e^{\lambda S_{(0,0),\mathbf{i}}}].$$

In this notation the defining equation for Λ (3.3) reads as

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{\Lambda_{(n,n)}(\lambda)}{||(n,n)||_1/2} = \sup_{n>0} \frac{\Lambda_{(n,n)}(\lambda)}{||(n,n)||_1/2}.$$

This can be extended to the statement that for all $\mathbf{i} \in \mathbb{L}_V$ we have

$$\frac{\Lambda_{\mathbf{i}}(\lambda)}{||\mathbf{i}||_1/2} \leq \Lambda(\lambda). \tag{4.9}$$

Indeed, if we define $\bar{\mathbf{i}}$ as the vertex one gets by interchanging the coordinates of $\mathbf{i}$, then $\mathbf{i} + \bar{\mathbf{i}}$ is a vertex on I_1 with $||\mathbf{i} + \bar{\mathbf{i}}||_1 = 2||\mathbf{i}||_1$. Thus we have

$$\begin{aligned}
\Lambda_{||\mathbf{i}||_1}(\lambda) &= \log \mathbb{E}[e^{\lambda S_{(0,0),\mathbf{i}+\bar{\mathbf{i}}}}] \\
&\geq \log \mathbb{E}[e^{\lambda(S_{(0,0),\mathbf{i}} + S_{\mathbf{i},\mathbf{i}+\bar{\mathbf{i}}})}] \\
&= 2 \log \mathbb{E}[e^{\lambda S_{(0,0),\mathbf{i}}}] \\
&= 2\Lambda_{\mathbf{i}}(\lambda)
\end{aligned}$$

by superadditivity and independence. This proves (4.9).

From this point on everything proceeds as in Section 3.3: The analogues to Lemmas 2 and 3 can be proven, especially the basic equation (3.19) holds in the adapted formulation and we also have

$$\mathbb{P}(S_{(0,0),\mathbf{i}} > t) \leq e^{-t\theta^*}$$

for all $\mathbf{i} \in \mathbb{L}_V$. The only thing that changes is the number of summands in the analogue of the first sum in (3.22), which is of the order $O(t^2)$. □

The statement corresponding to Corollary 2 reads as

Corollary 3. *There exists some constant $c > 0$ such that for t large enough we have*

$$\mathbb{P}(G > t) \leq ct^2 e^{-t\theta^*}.$$

4.6 Large deviations for optimal local scores

From the results in Section 4.5 the following result immediately derives, which characterizes the large deviations of $M_{(0,0),(n,m)}$.

Theorem 8. *Let m, n and t tend to infinity in such a way that $t = o(\min(m,n))$ and $\log(mn) = o(t)$. Then*

$$\lim_{t\to\infty} -\frac{1}{t} \log \mathbb{P}(M_{(0,0),(m,n)} > t) = \theta^* \tag{4.10}$$

Proof. In analogy to (4.8) we define for every $\mathbf{i} \in \mathbb{L}_V$

$$G_{\mathbf{i}} := \max_{\mathbf{j} \geq \mathbf{i}} S_{\mathbf{i},\mathbf{j}}.$$

It is clear that

$$\mathbb{P}(M_{(0,0),(m,n)} > t) \leq \mathbb{P}\Big(\max_{(0,0) \leq \mathbf{i} \leq (m,n)} G_{\mathbf{i}} > t \Big) \leq mn \cdot \mathbb{P}(G > t). \tag{4.11}$$

Taking logs and dividing by $-t$ gives

$$\lim_{t \to \infty} -\frac{1}{t} \log \mathbb{P}(M_{(0,0),(m,n)} > t) \geq \theta^*$$

by Theorem 7 and since we assumed $\log(mn) = o(t)$.

For the reverse inequality we can simply repeat the second part of the proof of Lemma 3 on page 38. Observe that in the notation we introduced there we have

$$\mathbb{P}(T_{0,n_t^*} > t) \leq \mathbb{P}(M_{(0,0),(m,n)} > t),$$

since $t = o(\min(m,n))$. It follows

$$\lim_{t \to \infty} -\frac{1}{t} \log \mathbb{P}(M_{(0,0),(m,n)} > t) \leq \lim_{t \to \infty} -\frac{1}{t} \log \mathbb{P}(T_{0,n_t^*} > t) = \theta^*.$$

□

From (4.11) we can give an upper bound for $\mathbb{P}(M_{(0,0),(m,n)} > t)$ which is in the same spirit as Corollary 3.

Corollary 4. *There exists some constant $c > 0$ such that for t large enough we have*

$$\mathbb{P}(M_{(0,0),(m,n)} > t) \leq cmnt^2 e^{-t\theta^*}.$$

4.7 A phase transition for the optimal local score

The results we have so far can nicely be used to prove the phase transition result for optimal local scores that has first been proven by Arratia and Waterman in [AW94] and has then been refined by Zhang in [Zha95].

4.7.1 The logarithmic phase

The logarithmic phase is the one which is more relevant for practical issues. Heuristically said, in the logarithmic phase optimal global alignments are always very bad, so one can do substantially better when optimizing locally.

Theorem 9. *Let the scoring scheme be such that $\gamma < 0$ and such that there is a unique positive solution θ^* of $\Lambda(\lambda) = 0$. Then*

$$\lim_{n \to \infty} \frac{M_{0,n}}{\log n} = \frac{2}{\theta^*} \quad a.s.$$

The proof will be split into two parts — one for the upper bound and one for the lower bound. We put the two parts into two different lemmas, since they use completely different techniques. We start with the upper bound, which relies more on the results from the previous sections.

Lemma 9. *Under the assumptions of Theorem 9 we have*

$$\limsup_{n\to\infty} \frac{M_{0,n}}{\log n} \leq \frac{2}{\theta^*} \quad a.s.$$

Proof. Fix $\varepsilon > 0$. In analogy to (4.8) we define for every $\mathbf{i} \in \mathbb{L}_V$

$$G_{\mathbf{i}} := \max_{\mathbf{j}\geq\mathbf{i}} S_{\mathbf{i},\mathbf{j}}. \tag{4.12}$$

It is clear that for any t and any n

$$\begin{aligned}\mathbb{P}(M_{0,n} > t) &\leq \mathbb{P}\left(\bigcup_{(0,0)\leq\mathbf{i}\leq(n,n)} \{G_{\mathbf{i}} > t\}\right)\\ &\leq n^2\mathbb{P}(G > t)\\ &\leq ct^2n^2e^{-t\theta^*},\end{aligned}$$

where we used Corollary 3. Especially for $t = (2+\varepsilon)/\theta^* \log n$ we get

$$\begin{aligned}\mathbb{P}\left(M_{0,n} > \frac{2+\varepsilon}{\theta^*}\log n\right) &\leq c\left(\frac{2+\varepsilon}{\theta^*}\right)^2 (\log n)^2 n^2 e^{-(2+\varepsilon)\log n}\\ &= c\left(\frac{2+\varepsilon}{\theta^*}\right)^2 \frac{(\log n)^2}{n^\varepsilon} \xrightarrow[n\to\infty]{} 0.\end{aligned} \tag{4.13}$$

If we replace n in this estimate by 2^k and use the Borel-Cantelli lemma we immediately get that

$$\limsup_{k\to\infty} \frac{M_{0,2^k}}{\log 2^k} \leq \frac{2}{\theta^*} \quad \text{a.s.}$$

The funny thing is that, by the logarithmic scaling, this is enough to finish the proof. To see this, define $k(n)$ via

$$2^{k(n)-1} < n \leq 2^{k(n)}.$$

Then we have

$$\frac{M_{0,n}}{\log n} \leq \frac{M_{0,2^{k(n)}}}{\log n} = \frac{M_{0,2^{k(n)}}}{\log 2^{k(n)}} \cdot \frac{1}{1 + \log\left(\frac{n}{2^{k(n)}}\right)/\log 2^{k(n)}},$$

where the last fraction goes to 1 for $n \to \infty$, since $\log(n/2^{k(n)}) > \log(1/2)$. □

Remark 11. Since we have packed most of the work into Corollary 3, our proof of Lemma 9 is substantially shorter than the corresponding proof by Arratia and Waterman ([AW94], Lemma 2, p. 206). That one can derive an almost sure upper bound from the original estimate (4.13) has been remarked by Zhang in [Zha95], although he does not give the argument.

Let us now come to the lower bound

Lemma 10. *Under the assumptions of Theorem 9 we have*

$$\liminf_{n\to\infty} \frac{M_{0,n}}{\log n} \geq \frac{2}{\theta^*} \quad a.s.$$

Proof. We will not repeat the proof here, which can be found in [Zha95], but merely make a few remarks.

Observe that in our context the unique positive solutions θ_m^* of $\Lambda_m(\lambda) = 0$ converge from above to θ^*. Thus it is enough to pick an $\varepsilon > 0$ and choose m large enough, such that

$$\frac{2}{\theta_m^*} \geq \frac{2}{\theta^*} - \varepsilon.$$

It is clear that (in the notation of [Zha95])

$$\frac{M_{0,mn}}{\log n} \geq \frac{\mathcal{M}_n}{\log n} \xrightarrow[n\to\infty]{} \frac{2}{\theta_m^*} \quad \text{a.s.}$$

□

4.7.2 The linear phase

In the linear phase the optimal local alignments essentially look like the optimal global alignments. The following theorem states this.

Theorem 10. *Let the scoring scheme be such that $\gamma > 0$. Then*

$$\lim_{n\to\infty} \frac{M_{0,n}}{n} = \gamma \quad a.s.$$

The proof can essentially be found in [AW94] (Lemma 3, p. 209, together with the remark that follows the proof). Again, the large deviations result Theorem 1 gives a better guideline to organize the proof.

Observe first that clearly

$$M_{0,n} \geq S_{0,n}$$

and we get by (4.3)

$$\liminf_{n\to\infty} \frac{M_{0,n}}{n} \geq \gamma \quad \text{a.s.} \tag{4.14}$$

For the other direction we start with

Lemma 11. *For every $\varepsilon > 0$ there is a constant $\mu := \mu(\varepsilon) > 0$ such that*

$$\mathbb{P}(M_{0,n} > (1+\varepsilon)\gamma n) \leq n^4 e^{-\mu\gamma(1+\varepsilon)n}.$$

Proof. Recall that

$$M_{0,n} = M_{(0,0),(n,n)} := \max_{(0,0)\leq \mathbf{i}\leq \mathbf{j}\leq (n,n)} S_{\mathbf{i},\mathbf{j}},$$

so we are done if we can show for every $(0,0) \leq \mathbf{i} \leq (n,n)$ that

$$\mathbb{P}(S_{(0,0),\mathbf{i}} > (1+\varepsilon)\gamma n) \leq e^{-\mu\gamma(1+\varepsilon)n}. \tag{4.15}$$

As in the proof of Theorem 2 we rely on the graphical intuition that is motivated by two facts. First we have the basic inequality

$$\mathbb{P}(S_{(0,0),\mathbf{i}} > t) \leq e^{-t\mu_{t,\mathbf{i}}}, \tag{4.16}$$

where

$$\mu_{t,\mathbf{i}} := \lambda_{t,\mathbf{i}} - \tfrac{1}{||\mathbf{i}||_1/2}\Lambda_{\mathbf{i}}(\lambda_{t,\mathbf{i}})/q_{t,\mathbf{i}}$$

with

$$q_{t,\mathbf{i}} := \frac{t}{||\mathbf{i}||_1/2}$$

and $\lambda_{t,\mathbf{i}}$ is such that

$$\tfrac{1}{||\mathbf{i}||_1/2}\Lambda'_{\mathbf{i}}(\lambda_{t,\mathbf{i}}-) \leq q_{t,\mathbf{i}} \leq \tfrac{1}{||\mathbf{i}||_1/2}\Lambda'_{\mathbf{i}}(\lambda_{t,\mathbf{i}}+). \tag{4.17}$$

And second we have

$$\tfrac{1}{||\mathbf{i}||_1/2}\Lambda_{\mathbf{i}}(\lambda) \leq \Lambda(\lambda)$$

for all $\mathbf{i} \geq (0,0)$ and all $\lambda \geq 0$ and all $\Lambda_{\mathbf{i}}$'s are convex as well as Λ. (Observe that (4.16) is even valid, if t and $\mathbf{i}$ are such that (4.17) does not have a solution. In such cases we simply have $\mathbb{P}(S_{(0,0),\mathbf{i}} > t) = 0$.)

We now define μ. To do this, choose $\bar{\lambda} > 0$ as the unique value for which

$$\Lambda'(\bar{\lambda}-) \leq (1+\varepsilon)\gamma \leq \Lambda'(\bar{\lambda}+)$$

and set

$$\mu := \bar{\lambda} - \frac{\Lambda(\bar{\lambda})}{(1+\varepsilon)\gamma}.$$

If we now set $t := (1+\varepsilon)\gamma n$, it is clear that for all $(0,0) \leq \mathbf{i} \leq (n,n)$ we have

$$q_{t,\mathbf{i}} \geq (1+\varepsilon)\gamma$$

and from this one sees immediately that

$$\mu_{t,\mathbf{i}} \geq \mu$$

which proves (4.15). □

Lemma 11 together with (4.14) now give that

$$\lim_{n\to\infty} \frac{M_{0,n}}{n} = \gamma \quad \text{in probability.}$$

To get almost sure convergence an interesting new turn comes into the proof. First, since $M_{0,n}/n$ is bounded from above and below, we also have that

$$\lim_{n\to\infty} \frac{\mathbb{E}[M_{0,n}]}{n} = \gamma,$$

a fact which is not clear from superadditivity. The same program that has been carried out in Section 4.3 for $S_{0,n}$ can be repeated for $M_{0,n}$ to show that $M_{0,n}/n$ concentrates nicely around γ. A standard use of the Borel-Cantelli lemma then gives almost sure convergence and the proof of Theorem 10 is complete.

4.8 Rate of convergence to the growth constant

In this section we will give bounds for the rate of convergence of $\mathbb{E}[S_{0,n}]/n$ towards the growth constant γ introduced in Section 4.2. For this we will use the concentration results derived in Section 4.3. The main result goes as follows

Theorem 11. *There exists a constant c_4 such that for large enough n*

$$\mathbb{E}[S_{0,n}] \le n\gamma \le \mathbb{E}[S_{0,n}] + c_4(n\log n)^{1/2}. \tag{4.18}$$

Remark 12. A similar result has first been proved by Alexander ([Ale93]) for first-passage percolation. In [Ale94] he used his methods to prove the analogous result for the mean length of the longest common subsequence of two random sequences. A very nice proof of the latter result has then been given by Rhee in [Rhe95] (cf. also [Ste97], Section 1.6). Later in [Ale97] Alexander developed a more abstract and general technique which could also be used in our case. Using this technique one could extend the result to optimal global scores of sequences of *unequal* length (i.e. where the ratio of the sequence length converges to some constant different from 1).

Anyway, to make our exposition more readable we have chosen to follow Rhee's proof for our result.

To prove Theorem 11 we need the following Lemma which is the analogue of Lemma 1 in [Rhe95] or to Lemma 1.6.1 in [Ste97].

Lemma 12. *We have*

$$\mathbb{P}(S_{0,4n} \ge 2t) \le (4n)^4 \mathbb{P}(S_{0,2n} \ge t - 3c_2/2)^{1/2},$$

where c_2 is the constant from Lemma 4.

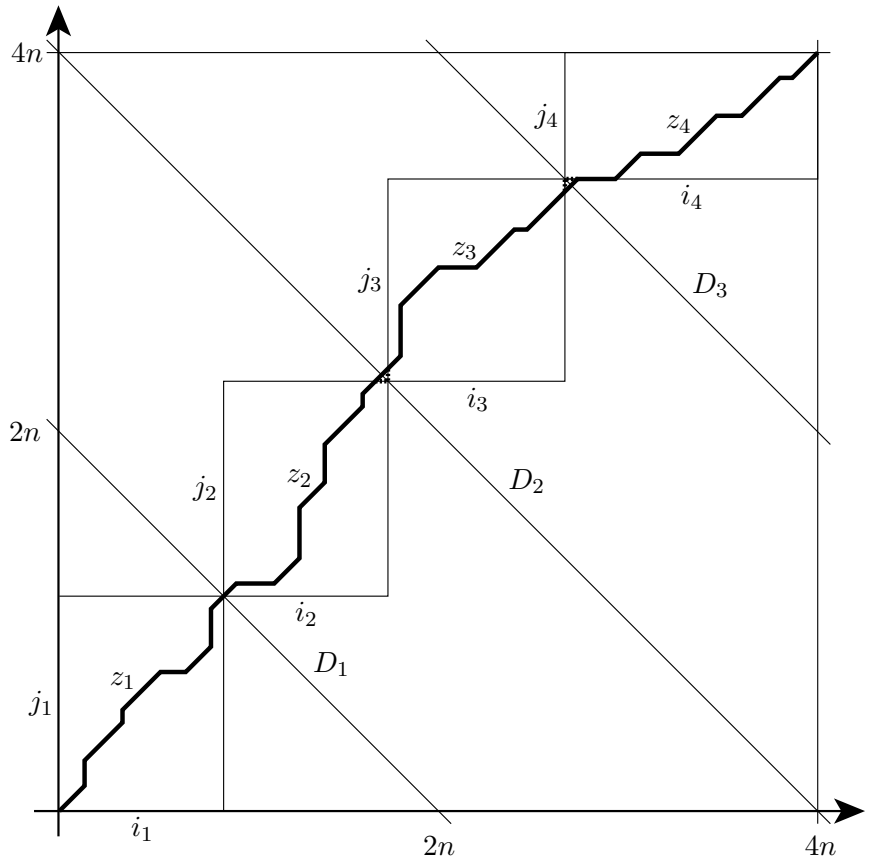

Figure 4.3: *Splitting $\tilde{z}$ into four parts.*

Proof. Let z be an optimal alignment for $S_{0,4n}$. There is a total number of $8n$ letters involved in the alignment. We can split z into four parts where in each part exactly $2n$ letters are used. Basically this can be done by cutting z at the points where it crosses the three diagonals D_1, D_2 and D_3 (cf. Figure 4.3). The only problem arising is that on some of the diagonals z might not hit any vertex on them. This is analogous to Lemma 4 (cf. also Figure 4.2). To make sure that such a path z can still be split, we have to modify it as usually. Denote this modified path by $\tilde{z}$. Whereas we have

$$s_z(\mathbf{X}_0^{4n}, \mathbf{Y}_0^{4n}) = S_{0,4n},$$

because of the optimality of z, we still have

$$s_{\tilde{z}}(\mathbf{X}_0^{4n}, \mathbf{Y}_0^{4n}) \geq S_{0,4n} - 3c_2.$$

Denote (cf. Figure 4.3) the four parts we obtain by splitting $\tilde{z}$ by z_l and the number of involved letters from the first resp. the second sequence by i_l resp. j_l, where $l = 1, \ldots, 4$. Observe that the construction is such that

$i_l + j_l = 2n$ for $l = 1, \ldots, 4$. Furthermore we denote by S_l, $l = 1, \ldots, 4$, the scores of the four parts, i.e. we have

$$s_{\tilde{z}}(\mathbf{X}_0^{4n}, \mathbf{Y}_0^{4n}) = S_1 + S_2 + S_3 + S_4.$$

Denote now by $\mathcal{I}$ the set of all subintervals $I_{\mathbf{X}}$ and $I_{\mathbf{Y}}$ of $\mathbf{X}_0^{4n}$ and $\mathbf{Y}_0^{4n}$ with $|I_{\mathbf{X}}| + |I_{\mathbf{Y}}| = 2n$, i.e.

$$\mathcal{I} = \{(I_{\mathbf{X}}, I_{\mathbf{Y}}) = (\mathbf{X}_i^j, \mathbf{Y}_k^l): \\ 0 \le i \le j \le 4n, 0 \le k \le l \le 4n, (j-i) + (l-k) = 2n\}.$$

From all this follows that

$$\begin{aligned}\{S_{0,4n} \ge 2t\} &\subseteq \{\max S_l \ge (2t - 3c_2)/4\} \\ &\subseteq \bigcup_{(I_{\mathbf{X}}, I_{\mathbf{Y}}) \in \mathcal{I}} \{s(I_{\mathbf{X}}, I_{\mathbf{Y}}) \ge (2t - 3c_2)/4\}\end{aligned}$$

and together with the crude upper bound $|\mathcal{I}| \le (4n)^4$ this implies

$$\mathbb{P}(S_{0,4n} \ge 2t) \le (4n)^4 \max_{(I_{\mathbf{X}}, I_{\mathbf{Y}}) \in \mathcal{I}} \mathbb{P}(s(I_{\mathbf{X}}, I_{\mathbf{Y}}) \ge (2t - 3c_2)/4).$$

This last maximum can be reexpressed by

$$\begin{aligned}\max_{(I_{\mathbf{X}}, I_{\mathbf{Y}}) \in \mathcal{I}} &\mathbb{P}(s(I_{\mathbf{X}}, I_{\mathbf{Y}}) \ge (2t - 3c_2)/4) = \\ &= \max_{i,j:\, i+j=2n} \mathbb{P}(s(\mathbf{X}_0^i, \mathbf{Y}_0^j) \ge (2t - 3c_2)/4).\end{aligned}$$

A simple symmetry argument now gives for arbitrary i, j with $i + j = 2n$

$$\begin{aligned}&\mathbb{P}(S_{0,2n} \ge t - 3c_2/2) \\ &\quad \ge \mathbb{P}(s(\mathbf{X}_0^i, \mathbf{Y}_0^j) \ge (2t - 3c_2)/4, s(\mathbf{X}_i^{2n}, \mathbf{Y}_j^{2n}) \ge (2t - 3c_2)/4) \\ &\quad = \mathbb{P}(s(\mathbf{X}_0^i, \mathbf{Y}_0^j) \ge (2t - 3c_2)/4)^2\end{aligned}$$

or

$$\max_{i,j:\, i+j=2n} \mathbb{P}(s(\mathbf{X}_0^i, \mathbf{Y}_0^j) \ge (2t - 3c_2)/4) \le \mathbb{P}(S_{0,2n} \ge t - 3c_2/2)^{1/2},$$

which completes the proof. □

Proof of Theorem 11. To prove the result we need the simple fact that for any real-valued random variable X and non-negative u we have

$$\begin{aligned}\mathbb{E}[X] &\le u + u\mathbb{P}(X \ge u) + \int_u^\infty \mathbb{P}(X \ge y)dy \\ &\le 2u + \int_u^\infty \mathbb{P}(X \ge y)dy.\end{aligned}$$

This can be proven using partial integration. We apply this to $X = S_{0,4n} - 2\mathbb{E}[S_{0,2n}]$ with $u = c(n \log n)^{1/2}$ for a constant c we choose later. This gives

$$\mathbb{E}[S_{0,4n}] - 2\mathbb{E}[S_{0,2n}] \leq \\ \leq 2c(n \log n)^{1/2} + \int_{c(n \log n)^{1/2}}^{\infty} \mathbb{P}(S_{0,4n} \geq y + 2\mathbb{E}[S_{0,2n}])dy.$$

Using Theorem 4 and Lemma 12 the integral can be bounded by

$$\begin{aligned} &\int_{c(n \log n)^{1/2}}^{\infty} \mathbb{P}(S_{0,4n} \geq y + 2\mathbb{E}[S_{0,2n}])dy \\ &\qquad \leq (4n)^4 \int_{c(n \log n)^{1/2}}^{\infty} \mathbb{P}(S_{0,2n} \geq y + \mathbb{E}[S_{0,2n}] - 3c_2/2)^{1/2} dy \\ &\qquad \leq 2(4n)^4 \int_{c(n \log n)^{1/2}}^{\infty} \exp\left(-\frac{(y - 3c_2/2)^2}{128 c_3^2 n}\right) dy. \end{aligned} \tag{4.19}$$

Some lines of elementary calculus, which we leave to the reader, show that for c large enough this whole term converges to 0 as $n \to \infty$. What we have proven so far, is

$$\frac{\mathbb{E}[S_{0,4n}]}{4n} - \frac{\mathbb{E}[S_{0,2n}]}{2n} \leq c\left(\frac{\log n}{n}\right)^{1/2}.$$

Replacing n by $2^k n$ we get

$$\frac{\mathbb{E}[S_{0,2^{k+2}n}]}{2^{k+2}n} - \frac{\mathbb{E}[S_{0,2^{k+1}n}]}{2^{k+1}n} \leq c\left(\frac{\log(2^k n)}{2^k n}\right)^{1/2}.$$

Summing this up over all $0 \leq k \leq j+1$ gives

$$\begin{aligned} \frac{\mathbb{E}[S_{0,2^{j+2}n}]}{2^{j+2}n} - \frac{\mathbb{E}[S_{0,2n}]}{2n} &= \sum_{k=0}^{j+1} \frac{\mathbb{E}[S_{0,2^{k+2}n}]}{2^{k+2}n} - \frac{\mathbb{E}[S_{0,2^{k+1}n}]}{2^{k+1}n} \\ &\leq c \sum_{k=0}^{j+1} \left(\frac{\log(2^k n)}{2^k n}\right)^{1/2} \\ &\leq c_4 \left(\frac{\log(2n)}{2n}\right)^{1/2} \end{aligned}$$

for c_4 large enough. Letting $j \to \infty$ proves the theorem for even n. The case of odd n trivially follows by monotonicity. □

4.9 Rate of convergence to Λ

As we have seen in several results, the unique positive zero θ^*, if it exists, of the function Λ from Theorem 1 is of great importance. Since

$$\Lambda(\lambda) = \lim_{n \to \infty} \frac{1}{n} \Lambda_n(\lambda) = \sup_{n > 0} \frac{1}{n} \Lambda_n(\lambda) \tag{4.20}$$

for all $\lambda \geq 0$ it is also the case that

$$\theta^* = \lim_{n\to\infty} \theta_n^* = \inf_{n>0} \theta_n^*,$$

where the θ_n^* are the unique positive zeros of the functions Λ_n, which exist for large enough n. We have the following result about the convergence in (4.20).

Theorem 12. *We have*

$$0 \leq \Lambda(\lambda) - \frac{1}{2n}\Lambda_{2n}(\lambda) \leq \frac{1}{n}[2\log(2) + 2\log(2n+1) + \lambda c_2],$$

where c_2 is the constant from Lemma 4.

In order to prove Theorem 12 we need some more notation and a couple of lemmas.

Let us start by defining

$$\hat{S}_{\mathbf{i},n} := \max_{\substack{\mathbf{j}\geq\mathbf{i} \\ ||\mathbf{j}-\mathbf{i}||_1=2n}} S_{\mathbf{i},\mathbf{j}}$$

and set

$$\hat{\Lambda}_n(\lambda) := \log \mathbb{E}[e^{\lambda \hat{S}_{(0,0),n}}].$$

We have

Lemma 13. *We have*

$$2\hat{\Lambda}_n(\lambda) - \log(2n+1) \leq \hat{\Lambda}_{2n}(\lambda) \leq 2\hat{\Lambda}_n(\lambda) + \log(2n+1) + \lambda c_2.$$

Proof. The optimal alignment for $\hat{S}_{(0,0),2n}$ can be split up in the same fashion as in Lemma 4 to get

$$\hat{S}_{(0,0),2n} - c_2 \leq \max_{\substack{\mathbf{i}\geq(0,0) \\ ||\mathbf{i}||_1=2n}} \{S_{(0,0),\mathbf{i}} + \hat{S}_{\mathbf{i},n}\} \leq S_{(0,0),2n}.$$

From this we first get for $\lambda \geq 0$

$$\begin{aligned} e^{-\lambda c_2} e^{\lambda \hat{S}_{(0,0),2n}} &\leq \max_{\substack{\mathbf{i}\geq(0,0) \\ ||\mathbf{i}||_1=2n}} e^{\lambda(S_{(0,0),\mathbf{i}} + \hat{S}_{\mathbf{i},n})} \\ &\leq \sum_{\substack{\mathbf{i}\geq(0,0) \\ ||\mathbf{i}||_1=2n}} e^{\lambda(S_{(0,0),\mathbf{i}} + \hat{S}_{\mathbf{i},n})}. \end{aligned}$$

This yields, using the independence of $S_{(0,0),\mathbf{i}}$ and $\hat{S}_{\mathbf{i},n}$ for fixed $\mathbf{i}$ and the fact that $\hat{S}_{\mathbf{i},n}$ has the same distribution as $\hat{S}_{(0,0),n}$,

$$\begin{aligned} e^{-\lambda c_2}\mathbb{E}[e^{\lambda \hat{S}_{(0,0),2n}}] &\leq \sum_{\substack{\mathbf{i}\geq(0,0)\\ ||\mathbf{i}||_1=2n}} \mathbb{E}[e^{\lambda(S_{(0,0),\mathbf{i}}+\hat{S}_{\mathbf{i},n})}] \\ &= \mathbb{E}[e^{\lambda \hat{S}_{(0,0),n}}] \sum_{\substack{\mathbf{i}\geq(0,0)\\ ||\mathbf{i}||_1=2n}} \mathbb{E}[e^{\lambda S_{(0,0),\mathbf{i}}}] \\ &\leq (2n+1)(\mathbb{E}[e^{\lambda \hat{S}_{(0,0),n}}])^2. \end{aligned} \tag{4.21}$$

On the other hand we have for every $\mathbf{i} \geq (0,0)$ with $||\mathbf{i}||_1 = 2n$,

$$e^{\lambda \hat{S}_{(0,0),2n}} \geq e^{\lambda(S_{(0,0),\mathbf{i}}+\hat{S}_{\mathbf{i},n})}$$

and as before it follows that

$$\mathbb{E}[e^{\lambda \hat{S}_{(0,0),2n}}] \geq \mathbb{E}[e^{\lambda \hat{S}_{(0,0),n}}]\mathbb{E}[e^{\lambda S_{(0,0),\mathbf{i}}}].$$

Summing this up over the $2n+1$ different $\mathbf{i}$'s for which this is valid gives

$$\begin{aligned} (2n+1)\mathbb{E}[e^{\lambda \hat{S}_{(0,0),2n}}] &\geq \mathbb{E}[e^{\lambda \hat{S}_{(0,0),n}}]\mathbb{E}\Big[\sum_{\substack{\mathbf{i}\geq(0,0)\\ ||\mathbf{i}||_1=2n}} e^{\lambda S_{(0,0),\mathbf{i}}}\Big] \\ &\geq (\mathbb{E}[e^{\lambda \hat{S}_{(0,0),n}}])^2. \end{aligned} \tag{4.22}$$

Taking logs in both (4.21) and (4.22) gives the result. □

The next lemma shows how Λ and $\hat{\Lambda}$ can be compared.

Lemma 14.

$$\begin{aligned} \hat{\Lambda}_{2n}(\lambda) \geq \Lambda_{2n}(\lambda) &\geq 2\hat{\Lambda}_n(\lambda) - 2\log(2n+1) \\ &\geq \hat{\Lambda}_{2n}(\lambda) - 3\log(2n+1) - \lambda c_2. \end{aligned}$$

Proof. Only the second inequality requires some work, the first is obvious and the third follows from Lemma 13.

Use the right-hand side of Lemma 4 to get

$$S_{0,2n} \geq \max_{\substack{\mathbf{i}\geq(0,0)\\ ||\mathbf{i}||_1=2n}} \{S_{(0,0),\mathbf{i}} + S_{\mathbf{i},(2n,2n)}\}.$$

By symmetry for every $\mathbf{j} \geq (0,0)$ with $||\mathbf{j}||_1 = 2n$

$$\begin{aligned} \mathbb{E}[e^{\lambda S_{0,2n}}] &\geq \mathbb{E}\Big[\max_{\substack{\mathbf{i}\geq(0,0)\\ ||\mathbf{i}||_1=2n}} e^{\lambda(S_{(0,0),\mathbf{i}}+S_{\mathbf{i},(2n,2n)})}\Big] \\ &\geq \mathbb{E}[e^{\lambda(S_{(0,0),\mathbf{j}}+S_{\mathbf{j},(2n,2n)})}] \\ &= (\mathbb{E}[e^{\lambda S_{(0,0),\mathbf{j}}}])^2. \end{aligned} \tag{4.23}$$

A simple calculation gives

$$(2n+1)\max_{\substack{\mathbf{j}\geq(0,0)\\ ||\mathbf{j}||_1=2n}}\mathbb{E}[e^{\lambda S_{(0,0),\mathbf{j}}}]\geq\sum_{\substack{\mathbf{j}\geq(0,0)\\ ||\mathbf{j}||_1=2n}}\mathbb{E}[e^{\lambda S_{(0,0),\mathbf{j}}}]$$
$$\geq\mathbb{E}[e^{\lambda\hat{S}_{(0,0),n}}].$$

This, together with (4.23), gives

$$\mathbb{E}[e^{\lambda S_{0,2n}}]\geq\max_{\substack{\mathbf{j}\geq(0,0)\\ ||\mathbf{j}||_1=2n}}(\mathbb{E}[e^{\lambda S_{(0,0),\mathbf{j}}}])^2=\Big(\max_{\substack{\mathbf{j}\geq(0,0)\\ ||\mathbf{j}||_1=2n}}\mathbb{E}[e^{\lambda S_{(0,0),\mathbf{j}}}]\Big)^2$$
$$\geq\frac{1}{(2n+1)^2}(\mathbb{E}[e^{\lambda\hat{S}_{(0,0),n}}])^2.$$

□

Lemma 15. $\lim_{n\to\infty}\hat{\Lambda}_n(\lambda)/n=\Lambda(\lambda)$ *and*

$$-[4\log(2)+\log(n+1)]\leq n\Lambda(\lambda)-\hat{\Lambda}_n(\lambda)\leq 4\log(2)+\log(n+1)+\lambda c_2 \quad (4.24)$$

Proof. It is clear from Lemma 14 that $\hat{\Lambda}_n(\lambda)/n$ converges to $\Lambda(\lambda)$ through even n's. Also, $|\hat{S}_{2n}-\hat{S}_{2n+1}|\leq c$ for some constant c which depends on the scoring scheme, but not on n. The convergence thus follows.

For (4.24) use Lemma 13 and Lemma 18 from the Appendix with $l(n)=\log(n+1)$ and $u(n)=\log(n+1)+\lambda c_2$ and observe that

$$\sum_{i=1}^{\infty}2^{-i}\log(2^in+1)\leq 4\log 2+\log(n+1).$$

□

Proof of Theorem 12. By Lemma 14 we can compare Λ_n and $\hat{\Lambda}_n$ and use Lemma 15 to get

$$\begin{aligned}0&\leq\Lambda(\lambda)-\frac{1}{2n}\Lambda_{2n}(\lambda)\\&=\Lambda(\lambda)-\frac{1}{2n}\hat{\Lambda}_{2n}(\lambda)+\frac{1}{2n}[\hat{\Lambda}_{2n}(\lambda)-\Lambda_{2n}(\lambda)]\\&\leq\frac{1}{2n}[4\log(2)+\log(2n+1)+\lambda c_2]+\frac{1}{2n}[3\log(2n+1)+\lambda c_2]\\&=\frac{1}{n}[2\log(2)+2\log(2n+1)+\lambda c_2].\end{aligned}$$

□

4.10 Multiple splitting

We will now describe another way of deriving a result similar to the statement in Theorem 12. In spite of the disadvantage of being more complicated, this approach has the advantage of improving Theorem 12 and of providing some formulas which might be helpful for further investigations. We will start with a result which is much in the spirit of Lemma 4.

Let $m, n \in \mathbb{N}$ where n is a multiple of m. For some i with $-n \leq i \leq n$ we want to split the optimal alignment from $(0,0)$ to the point $(n+i, n-i)$ at the secondary diagonals $(D_j)_{1 \leq j \leq n/m-1}$, where

$$D_j := \{\mathbf{k} \colon (0,0) \leq \mathbf{k} \leq (n+i, n-i), ||\mathbf{k}||_1 = 2jm\}.$$

To each increasing path $\vec{\mathbf{k}} = (\mathbf{k}_0, \ldots, \mathbf{k}_{n/m})$, where

$$(0,0) = \mathbf{k}_0 \leq \mathbf{k}_1 \leq \ldots \leq \mathbf{k}_{n/m} = (n+i, n-i)$$

we assign a score $S_{\vec{\mathbf{k}}}$ via

$$S_{\vec{\mathbf{k}}} := \sum_{j=1}^{n/m} S_{\mathbf{k}_{j-1}, \mathbf{k}_j}$$

and denote the set of all such paths by $K_{n,m,i}$. It is clear that for any $\vec{\mathbf{k}} \in K_{n,m,i}$ we have

$$S_{(0,0),(n+i,n-i)} \geq S_{\vec{\mathbf{k}}}$$

which extends by maximization to

$$\max_{\vec{\mathbf{k}} \in K_{n,m,i}} S_{\vec{\mathbf{k}}} \leq S_{(0,0),(n+i,n-i)}. \tag{4.25}$$

On the other hand considerations similar to the ones used in Lemma 4 give that

$$S_{(0,0),(n+i,n-i)} \leq \max_{\vec{\mathbf{k}} \in K_{n,m,i}} S_{\vec{\mathbf{k}}} + c_2 n/m. \tag{4.26}$$

4.10.1 A relation between LMGFs

We now use the splitting to derive a relation between logarithmic moment generating functions. To do so we start by defining the logarithmic moment generating function of $S_{\vec{\mathbf{k}}}$ for $\vec{\mathbf{k}} \in K_{n,m,i}$ as

$$\Lambda_{\vec{\mathbf{k}}}(\lambda) := \log \mathbb{E}[e^{\lambda S_{\vec{\mathbf{k}}}}] = \sum_{j=1}^{n/m} \Lambda_{\mathbf{k}_j - \mathbf{k}_{j-1}}(\lambda). \tag{4.27}$$

This logarithmic moment generating function can now be conveniently used to describe exponential tilting along the path $\vec{\mathbf{k}}$. For a path $\vec{\mathbf{k}} \in K_{n,m,i}$ and $\lambda \in \mathbb{R}$ we define a measure $\mathbb{P}_{\vec{\mathbf{k}},\lambda}$ on $\Omega = \mathcal{A}^{n+i} \times \mathcal{A}^{n-i}$ by

$$\frac{\mathrm{d}\mathbb{P}_{\vec{\mathbf{k}},\lambda}}{\mathrm{d}\mathbb{P}}(\omega) := e^{\lambda S_{\vec{\mathbf{k}}}(\omega) - \Lambda_{\vec{\mathbf{k}}}(\lambda)}.$$

We use this tilting to get

$$\begin{aligned} e^{\Lambda_{(n+i,n-i)}(\lambda)} &= \mathbb{E}\left[e^{\lambda S_{(0,0),(n+i,n-i)}}\right] \\ &= \mathbb{E}\left[\frac{\sum_{\vec{\mathbf{k}} \in K_{n,m,i}} e^{\lambda S_{\vec{\mathbf{k}}}}}{\sum_{\vec{\mathbf{l}} \in K_{n,m,i}} e^{\lambda S_{\vec{\mathbf{l}}}}} \cdot e^{\lambda S_{(0,0),(n+i,n-i)}}\right] \\ &= \sum_{\vec{\mathbf{k}} \in K_{n,m,i}} e^{\Lambda_{\vec{\mathbf{k}}}(\lambda)} \mathbb{E}_{\vec{\mathbf{k}},\lambda}\left[\frac{e^{\lambda S_{(0,0),(n+i,n-i)}}}{\sum_{\vec{\mathbf{l}} \in K_{n,m,i}} e^{\lambda S_{\vec{\mathbf{l}}}}}\right]. \end{aligned}$$

Using (4.25) and (4.26) we get the final inequalities

$$e^{\Lambda_{(n+i,n-i)}(\lambda)} \geq \sum_{\vec{\mathbf{k}} \in K_{n,m,i}} e^{\Lambda_{\vec{\mathbf{k}}}(\lambda)} \mathbb{E}_{\vec{\mathbf{k}},\lambda}\left[\frac{\max_{\vec{\mathbf{l}}' \in K_{n,m,i}} e^{\lambda S_{\vec{\mathbf{l}}'}}}{\sum_{\vec{\mathbf{l}} \in K_{n,m,i}} e^{\lambda S_{\vec{\mathbf{l}}}}}\right] \tag{4.28}$$

and

$$e^{\Lambda_{(n+i,n-i)}(\lambda)} \leq e^{\lambda c_2 n/m} \sum_{\vec{\mathbf{k}} \in K_{n,m,i}} e^{\Lambda_{\vec{\mathbf{k}}}(\lambda)} \mathbb{E}_{\vec{\mathbf{k}},\lambda}\left[\frac{\max_{\vec{\mathbf{l}}' \in K_{n,m,i}} e^{\lambda S_{\vec{\mathbf{l}}'}}}{\sum_{\vec{\mathbf{l}} \in K_{n,m,i}} e^{\lambda S_{\vec{\mathbf{l}}}}}\right] \tag{4.29}$$

4.10.2 Exploiting the upper bound

Since the Expectation in (4.29) is bounded by one from above and by (4.27) we get

$$e^{\Lambda_{(n+i,n-i)}(\lambda)} \leq e^{\lambda c_2 n/m} \sum_{\vec{\mathbf{k}} \in K_{n,m,i}} e^{\sum_{j=1}^{n/m} \Lambda_{\mathbf{k}_j - \mathbf{k}_{j-1}}(\lambda)}. \tag{4.30}$$

We now change the parametrization in this formula to get a more convenient version. Observe that for each $\vec{\mathbf{k}}$ all the points $\mathbf{k}_j - \mathbf{k}_{j-1}$ are in the set

$$\{(m+h, m-h) \colon -m \leq h \leq m\}.$$

For a fixed $\vec{\mathbf{k}}$ and $-m \leq h \leq m$ we define

$$b_h := \#\{1 \leq j \leq n/m \colon \mathbf{k}_j - \mathbf{k}_{j-1} = (m+h, m-h)\}.$$

This defines a mapping from $K_{n,m,i}$ to the set

$$B_{n,m,i} := \left\{ b = (b_{-m}, \ldots, b_m) \colon \sum_{h=-m}^{m} b_h = n/m, \sum_{h=-m}^{m} h b_h = i \right\}.$$

This mapping is not one-to-one. In fact for a given $b \in B_{n,m,i}$ the number of $\vec{\mathbf{k}} \in K_{n,m,i}$ which are mapped onto it is easily calculated as

$$\binom{n/m}{b_{-m}, \ldots, b_m}$$

and for each such $\vec{\mathbf{k}}$ we have

$$\sum_{j=1}^{n/m} \Lambda_{\mathbf{k}_j - \mathbf{k}_{j-1}}(\lambda) = \sum_{h=-m}^{m} b_h \Lambda_{(m+h,m-h)}(\lambda).$$

In this parametrization the inequality (4.30) writes as

$$e^{\Lambda_{(n+i,n-i)}(\lambda)} \leq e^{\lambda c_2 n/m} \sum_{b \in B_{n,m,i}} \binom{n/m}{b_{-m}, \ldots, b_m} \exp\left(\sum_{h=-m}^{m} b_h \Lambda_{(m+h,m-h)}(\lambda) \right). \tag{4.31}$$

We want to use this inequality in the following way.

For the left hand side some limiting procedure can be carried out as follows. First we define $\rho := i/n$, so that we can write $n(1+\rho, 1-\rho)$ instead of $(n+i, n-i)$. Then we replace the n by kn, take logarithms and divide by kn. From superadditivity arguments it follows then that

$$\Lambda^{(\rho)}(\lambda) := \lim_{k \to \infty} \frac{1}{kn} \Lambda_{kn(1+\rho,1-\rho)}(\lambda)$$

exists as a convex function in λ. Of course $\Lambda^{(0)}$ is just the limiting logarithmic moment generating function Λ introduced above.

We now want to carry out the same kind of limiting procedure on the right-hand side.

To do this we need some more notation. First, to shorten notation we write

$$x_h := \Lambda_{(m+h,m-h)}(\lambda), \quad -m \leq h \leq m.$$

For a vector $x = (x_{-m}, \ldots, x_m) \in \mathbb{R}^{2m+1}$ define the function $f_x \colon \mathbb{R}^{2m+1} \to \mathbb{R}$ as

$$f_x(\pi_{-m}, \ldots, \pi_m) := -\sum_{h=-m}^{m} \pi_h [\log(\pi_h) - x_h].$$

Also we need

$$\tilde{f}_\rho(x) := \max\{f_x(\pi) \colon \pi \in P_{m,\rho}\},$$

where the maximization is over the set

$$P_{m,\rho} := \left\{ \pi = (\pi_{-m}, \ldots, \pi_m) \in [0,1]^{2m+1} \colon \sum_{h=-m}^{m} \pi_h = 1, \sum_{h=-m}^{m} h\pi_h = \rho m \right\}.$$

Approximating the sum in (4.31) by an integral and using a Laplace approximation we get the following Lemma which we will prove in Appendix A.5

Lemma 16. *We have*

$$\sum_{b\in B_{kn,m,kn\rho}} \binom{kn/m}{b_{-m},\ldots,b_m} \exp\left(\sum_{h=-m}^{m} b_h x_h\right) \leq$$
$$\leq C'' \left(\frac{kn}{m}\right)^{2m-\frac{1}{2}} e^{\frac{kn}{m}\tilde{f}_\rho(x)} \left\{1 + O\left(\left(\frac{kn}{m}\right)^{-1}\right)\right\}.$$

Taking logarithms, dividing by kn and letting $k \to \infty$ then gives the following upper bound for $\Lambda^{(\rho)}$

$$\Lambda^{(\rho)}(\lambda) \leq \frac{1}{m}\lambda c_2 + \frac{1}{m}\tilde{f}_\rho(x).$$

All that remains is to plug in a more concrete formula for $\tilde{f}_\rho(x)$. This can be done as shown in Appendix A.5. Especially it is shown there how to determine the value of λ_ρ which will be needed in the following. Recall that the vector x is symmetric, since its entries are defined as

$$x_h = \Lambda_{m+h,m-h}(\lambda)$$

so that $\lambda_0 = 0$. The final upper bound for $\Lambda^{(\rho)}$ now reads as

$$\Lambda^{(\rho)}(\lambda) \leq \frac{1}{m}\lambda c_2 + \frac{1}{m}\log\left(\sum_{h=-m}^{m} e^{\Lambda_{m+h,m-h}(\lambda)+\lambda_\rho h}\right) - \rho\lambda_\rho.$$

Especially for $\rho = 0$ we get the following corollary, which improves Theorem 12.

Corollary 5.

$$\Lambda(\lambda) - \frac{1}{m}\Lambda_m(\lambda) \leq \frac{1}{m}\lambda c_2 + \frac{1}{m}\log(2m+1).$$

Proof. Since for all $-m \leq h \leq m$

$$\Lambda_m(\lambda) := \Lambda_{m,m}(\lambda) \geq \Lambda_{m+h,m-h}(\lambda),$$

we get

$$\Lambda(\lambda) - \frac{1}{m}\Lambda_m(\lambda) \leq \frac{1}{m}\lambda c_2 + +\frac{1}{m}\log\left(\sum_{h=-m}^{m} \underbrace{e^{\Lambda_{m+h,m-h}(\lambda)-\Lambda_m(\lambda)}}_{\leq 1}\right)$$
$$\leq \frac{1}{m}\lambda c_2 + \frac{1}{m}\log(2m+1).$$

□

Chapter 5

Conclusions and outlook

We will now discuss in how far the results we derived in this thesis give insight into the asymptotic behaviour of the tail of the optimal local score's distribution. On the level of large deviations, our description of the tail's behaviour is quite complete. Still, it is not completely trivial to calculate the rate θ^* in practice. We will come back to this problem in Subsection 5.2.1.

Now, we want to discuss how one has to read our results in order to get finer statements about the behaviour of the tail. To better understand the context of this discussion we start by explaining the corresponding results for some simpler cases.

5.1 Finer tail asymptotics

The θ^* we characterized in Section 2.1 only answered the crudest of all questions about the optimal local score's tail behaviour: What is the leading term in its exponential decay? Immediately the question about a finer description of its asymptotic behaviour arises, meaning that ideally we would like to find two functions f and g for which

$$\mathbb{P}(M_{0,n} > t) \sim f(n)g(t)e^{-\theta^* t}$$

as both n and t tend to infinity.

In this section we will use the notation from Section 2.1.

5.1.1 The gapless case

In the gapless case the right asymptotics follow from (2.1), which gives that we should choose $f(n) = \eta n^2$ and $g(t) = C^*$, where $\eta C^* = K^*$. Without going much into detail, this holds, since $M_{0,n}$ can be interpreted as the maximum of ηn^2 asymptotically independent random variables, the $T_{\mathbf{i}}$'s for $\mathbf{i} \in [0, n]^2$, and where each of them has a very small chance of about $C^* e^{-\theta^* t}$

of exceeding the score t. To prove that

$$\mathbb{P}(T_{\mathbf{i}} > t) \sim C^* e^{-\theta^* t} \tag{5.1}$$

of course requires more elaborate methods than the ones we used to characterize θ^*. The main ingredient is renewal theory, details can be found in [Igl72] or [KD92]. Again, the rigorous result is a little bit more complicated, since the constant C^* can not be taken literally in the case where the possible values of the process $(S_{0,n})_{n\geq 0}$ do not give all of $\mathbb{R}$ but only a discrete subset. In this case, which is usually called the *lattice case*, the asymptotic behaviour in (5.1) has actually to be replaced by a lower and an upper bound with the constant C^* replaced by two constants.

Dembo, Karlin and Zeitouni use the Chen-Stein-method (cf. [AGlG89]) to prove a Poisson approximation and to get the asymptotic behaviour of $\mathbb{P}(M_{0,n} > t)$. This means that basically $\{T_{\mathbf{i}} \geq t\}$ and $\{T_{\mathbf{j}} \geq t\}$ are asymptotically independent for large t, if the two vertices $\mathbf{i}$ and $\mathbf{j}$ are far enough away from each other. Of course, in the course of the proof, the dependences between $\{T_{\mathbf{i}} \geq t\}$ and $\{T_{\mathbf{j}} \geq t\}$ have to be controlled — a task which, for *well-behaving* scoring schemes, can be done using large deviations arguments. In [DKZ94b], exact criteria are given which characterize the class of well-behaving scoring schemes — the scoring schemes which are commonly used in practice being amongst them.

5.1.2 An intermediate case

In [SY00] D. Siegmund and B. Yakir proved a tail asymptotics for the optimal local score in the somewhat artificial situation where gaps are allowed, but not too many of them. To be more precise, they restrict the overall number of horizontal and vertical stretches in the optimal alignment paths, whereas the overall length of those stretches is not confined. This in fact leads to a model which is settled between the gapless and the gapped alignments. What basically happens in this model is that it tries to find good gapless parts which are not too far away from each other and combines them to get a gapped alignment.

Concerning the asymptotic behaviour, it is of the same form as in the gapless case, only the value of the constant K^* has to be changed. Especially, the rate of decay θ^* is the same as in the gapless case, i.e. it can be characterized as the unique positive zero of the logarithmic moment generating function for the gapless case which we introduced in (2.4). As shown in Subsection 2.1.2, the rate of decay in the *real* gapped case, i.e. where all possible alignments are allowed, can be characterized as the unique positive zero of some limiting logarithmic moment generating function. It is defined in (2.10) and in nontrivial cases it is definitely *not identical* with the one in the gapless case. This shows that the restricted optimal score, which has been considered by Siegmund and Yakir, *cannot* serve as an approximation

for the full optimal local score. Already on the coarsest scale the tail of the latter behaves differently.

The main work in [SY00] goes into the problem of characterizing the changed constant in front of the exponential term. The dependences between different optimal alignments become much more complicated than in the pure gapless case and intricate calculations are necessary to control them. The constant depends on quite a complicated way on the chosen gap penalties, but formulas are given, which admit numerical calculations. Details related to the numerics can be found in [SS01]. For a simpler model that explains well the ideas behind Siegmund/Yakir's work cf. [MGW02].

5.1.3 The gapped case

In the gapped case a rigorous version of full tail asymptotics is still far away. It is conjectured, however, that the tail behaves much the same as in the gapless case, with different constants, of course.

But let us look at the results we have. An upper bound for the tail of $M_{0,n}$ is given by Corollary 4 which states that

$$\mathbb{P}(M_{(0,0),(m,n)} > t) \leq cmnt^2 e^{-t\theta^*} \tag{5.2}$$

for some constant c. What we would like to have next is some corresponding lower bound which would help to judge in how far this upper bound is sharp. Observe that to derive (5.2) it was not necessary to control the dependences between the $T_{\mathbf{i}}$'s. However, this would be necessary for a lower bound, i.e. if we were able to

(i) give a lower bound for the tail of the $T_{\mathbf{i}}$'s and

(ii) control the dependences between the $T_{\mathbf{i}}$'s,

this would lead to a lower bound for $\mathbb{P}(M_{(0,0),(m,n)} > t)$.

Let us discuss the first point in more detail. Recall the upper bound for $\mathbb{P}(T > t)$ which we gave in Corollary 3 and which states that

$$\mathbb{P}(T > t) \leq ct^2 e^{-t\theta^*}$$

for some constant c. We do not believe this upper bound to be sharp, but what is interesting is that, apart from the leading exponential term $e^{-t\theta^*}$, there is no other exponential term in t of some lower order, but only a polynomial one. A first step into the direction of a more refined result would be to also give a lower bound in which no exponential terms, apart from the leading one, would appear. This would then imply that also in the true asymptotic behaviour of $\mathbb{P}(T > t)$ there would appear no other exponential term apart from $e^{-t\theta^*}$.

The most promising way to get a lower bound is probably to take equation (2.8) and use it for $n = i^*$, where

$$i^* := i^*(t) := t/\Lambda'(\theta^*). \tag{5.3}$$

First observe that, although (2.12) is formulated in terms of logarithmic equivalence, it is in fact the case that a *polynomial correction term* suffices to relate the asymptotic behaviour of $\mathbb{P}(S_{i^*} > t)$ to the one of $\mathbb{P}(T > t)$. Then the rewriting of (2.8) gives (with $q^* := \Lambda'(\theta^*)$)

$$\mathbb{P}_0(S_{0,i^*} > i^*q^*) = e^{-i^*[\lambda_{i^*,q^*}q^* - (1/i^*)\Lambda_{i^*}(\lambda_{i^*,q^*})]} \times \\ \times \mathbb{E}_{\lambda_{i^*,q^*},i^*}[\exp(-\lambda_{i^*,q^*}(S_{0,i^*} - i^*q^*)); S_{0,i^*} > i^*q^*].$$

The exponential term on the right of this equation can be easily shown to be of the order $e^{-\theta^* t + O(\log t)} = \Omega(t^{-\alpha}) \cdot e^{-\theta^* t}$ for some $\alpha \geq 0$, since we have the convergence rate result

$$\Lambda(\lambda) - \frac{1}{n}\Lambda_n(\lambda) = O\left(\frac{\log n}{n}\right).$$

The problem is the remaining expectation, which is some kind of truncated moment generating function. As we already mentioned in Subsection 2.1.1, in the gapless case it is of the order $\Theta(t^{-1/2})$. This stems from the fact that in this case S_{0,i^*} is a sum of i^* i.i.d. random variables and thus the probability that it falls into the small strip $[\mathbb{E}[S_{0,i^*}], \mathbb{E}[S_{0,i^*}] + \delta]$ is of the order $\Theta((i^*)^{-1/2})$ and i^* can be replaced by t because of (5.3).

In the gapped case we do not have such a result. The only thing we know from the concentration inequalities is that, asymptotically, the total mass of the distribution of S_{0,i^*} is concentrated in a an interval around $\mathbb{E}[S_{0,i^*}]$ which has a length of the order $o(t^{1/2+\varepsilon})$ for any $\varepsilon > 0$. But does this imply that a fraction of the mass which is of order $\Omega(t^{-\beta})$ for some $\beta > 0$ can be found in the direct vicinity of $\mathbb{E}[S_{0,i^*}]$? This would then be sufficient to show that the whole expectation is of the order $\Omega(t^{-\beta})$ and finally we would get that

$$\mathbb{P}_0(T > t) \geq \Omega(t^{-\eta}) \cdot e^{-\theta^* t}$$

for some $\eta > 0$.

5.2 Assessing the statistical significance in practice

It would certainly go beyond the scope of this thesis to give a detailed overview of the approaches used in practice to assess the statistical significance of the results of sequence similarity searches. We will just explain the basic concepts in order to be able to describe in how far our result is useful.

Motivated by the rigorous results in the gapless case, which we sketched in Subsections 2.1.1 and 5.1.1, almost all authors assume that, also in the gapped case, the asymptotic distribution function of the optimal local score obtained by comparing two random sequences, is of the Gumbel-type

$$\lim_{n\to\infty} \mathbb{P}\left(M_{0,n} - \frac{2\log n}{\theta^*}\right) = \exp(-K^* \exp(-\theta^* t)). \tag{5.4}$$

First heuristical arguments and an empirical verification for this have been given by M. Waterman and M. Vingron in [WV94a] and [WV94b]. The heuristical argument given there is based on the Chen-Stein-method ([AGlG89]) but, in contrary to the gapless case, lacks rigour. Actually, Waterman and Vingron used computer simulations to check the validity of the heuristics.

Assuming the correctness of the Gumbel-form for the limiting distribution, the practitioner is left with mainly two problems.

(i) To estimate the two parameters K^* and θ^*.

(ii) To check how good the approximation by the limiting distribution function really is, since in practice we are only dealing with finite databases consisting of finite sequences.

This thesis can only make a substantial contribution to the first of these two problems and we will explain this in the next subsection. For the second we refer to [Pea98], [SV98], [Mot02] and the references therein.

5.2.1 Calculating the parameters

One of the main results of this thesis is that, if one assumes the correctness of the Gumbel-form, we have proved a characterization of the parameter θ^*, which in a natural way extends the gapless case. Whereas in the gapless case θ^* is given by the unique positive zero of the logarithmic moment generating function of the score of a randomly chosen letter pair, it is in the general case given by the unique positive zero of some *limiting* logarithmic moment generating function that arises from optimal global scores. Unfortunately, this limiting logarithmic moment generating function can *not* be calculated directly and neither can its unique positive zero; in practice it has to be estimated.

Our results provide a recipe how to do this. If we denote for each $n > 0$ by θ_n^* the unique positive zero of the function Λ_n, then it is clear from the convergence of $(1/n)\Lambda_n$ to Λ, that

$$\theta_n^* \searrow \theta^*$$

as $n \to \infty$. So, the basic idea is to estimate θ_n^* for large enough n in order to approximate θ^*.

Our result about the rate of convergence of the Λ_n's can be used here to get an estimate for the accuracy of the approximation. Observe that it follows from (2.15) that

$$\theta_n^* = \theta^* + O\left(\frac{\log n}{n}\right), \tag{5.5}$$

so it should be possible to control the error made in the approximation.

As we have learned after proving our results, the characterization of θ^* as the unique positive zero of some limiting logarithmic moment generating function had already been conjectured by R. Bundschuh in [Bun02]. Bundschuh gives heuristic arguments and also checks whether the estimation procedure for θ^* just described is useful in practice. His results concerning the latter issue are quite encouraging and we will describe them briefly.

Whereas we showed (5.5) rigorously, Bundschuh conjectures in [Bun02] that the convergence is of the form

$$\theta_n^* = \theta^* + \frac{c}{n} + O(n^{-2}), \tag{5.6}$$

where c is some constant, which might change from case to case. He then estimates θ_n^* for a number of moderately large values of n (concretely $n = 60$, $n = 80$ and $n = 100$) and fits a function of the form $a + b/n$ to this set of points. The fitted parameter a then serves as an estimator for θ^*. This is done for a number of scoring schemes and actually works well. The results were evaluated using independent, good estimates for θ^* which were obtained using extensive simulations.

Of course, in principle, it is as difficult to estimate the logarithmic moment generating function Λ_n as it is to estimate the tail of the distribution of $S_{0,n}$, but for the values of n used here this can be done in a very short time. The advantage is that the assumption made in (5.6) already works well for those small values of n.

Being now able to estimate the parameter θ^* in reasonable time, only the second parameter K^* remains to be estimated. Concerning this issue we do not have any comparable result, but observe two things.

(i) Although for practical purposes it might be sufficient to assume (5.4), it is by no means rigorously justified, that this is the right form, resp. the right scaling.

(ii) The parameter θ^* is by far the more important for assessing the statistical significance of database searches (cf. e.g. [Mot02]). Especially for very small p-values the influence of K^* becomes negligible.

Other approaches to estimate the parameters θ^* and K^* were mainly based on the use of extensive simulations (cf. [AG96], [WV94a], [WV94b], [OBH99]).

Appendix A

Appendix

A.1 Asymptotic relations

For convenience we use some notation for asymptotic relations which goes slightly beyond the usual Landau symbols. Our notation follows the one used in the analysis of computer algorithms (cf., e.g., [CLR90]). The classical Landau symbols are

$$f(x) = o(g(x)) \text{ as } x \to x_0 \quad :\Longleftrightarrow \quad \limsup_{x \to x_0} \frac{|f(x)|}{g(x)} = 0$$

and

$$f(x) = O(g(x)) \text{ as } x \to x_0 \quad :\Longleftrightarrow \quad \exists C > 0 \colon \limsup_{x \to x_0} \frac{|f(x)|}{g(x)} \leq C.$$

If one wants to give an asymptotic lower bound, instead of an upper bound, the following symbol is useful

$$f(x) = \Omega(g(x)) \text{ as } x \to x_0 \quad :\Longleftrightarrow \quad \exists C > 0 \colon \liminf_{x \to x_0} \frac{|f(x)|}{g(x)} \geq C,$$

which together with O gives a two-sided asymptotically sharp bound

$$f(x) = \Theta(g(x)) \text{ as } x \to x_0 \quad :\Longleftrightarrow \quad f(x) = O(g(x)) \wedge f(x) = \Omega(g(x)).$$

In large deviations theory one often is concerned with quantities which grow or decay exponentially fast. Sometimes one is only interested in the rate of such a growth or decay. It is therefore convenient to introduce (in the spirit of [dH00]) the notion of *logarithmic equivalence* as

$$f(x) \simeq g(x) \text{ as } x \to \infty \quad :\Longleftrightarrow \quad \lim_{x \to \infty} \frac{1}{x} (\log(f(x)) - \log(g(x))) = 0.$$

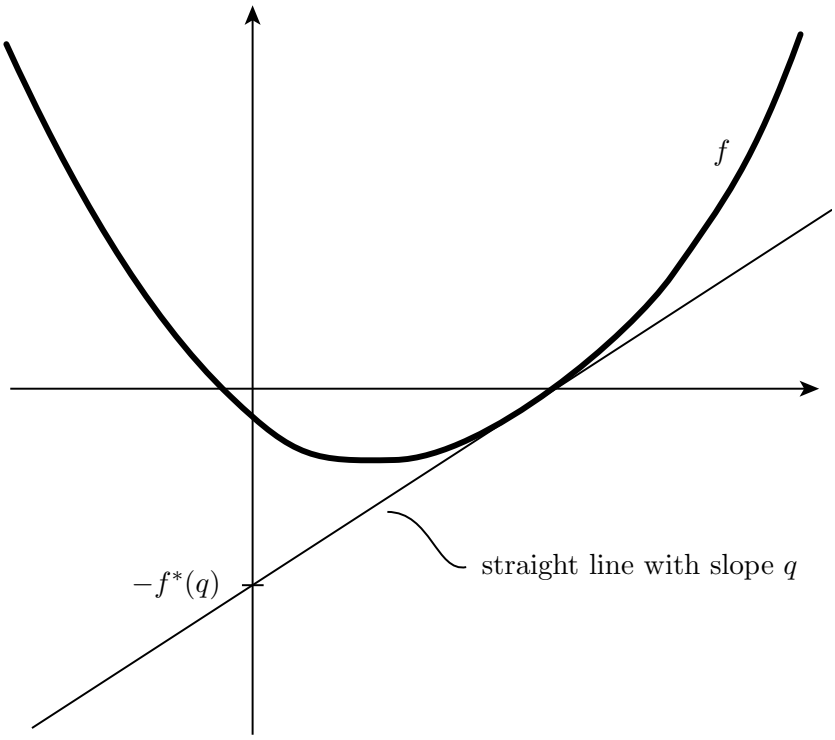

Figure A.1: *The geometrical intuition behind convex conjugation.*

Logarithmic equivalence has one especially important property, namely that

$$f(x) + g(x) \simeq \max(f(x), g(x)).$$

This is known as the "largest-exponent-wins" property. Of course it holds for all sums of finitely many functions as long as the number of summands does not grow exponentially in x.

A.2 Convex conjugation

Another principle, which plays a major role in many of the arguments of this thesis, is convex conjugation. We will give an informal and intuitive presentation of the basic ideas, for a more thorough threatment the reader should consult the Rockafellar's book ([Roc79]).

To any convex function f its *convex conjugate* f^* can be defined as

$$f^*(q) := \sup_{x \in \mathbb{R}} \{xq - f(x)\}. \tag{A.1}$$

We always think of convex functions as being defined on all of $\mathbb{R}$ and taking values in $\overline{\mathbb{R}} = [-\infty, \infty]$, since a convex function $f : D \to \mathbb{R}$ can always be

extended to a convex function in our sense by setting $f(x) := \infty$ for $x \notin D$. The set

$$\mathcal{D}_f := \{x \in \mathbb{R} \colon f(x) < \infty\}$$

is called the *domain* of f.

The very extreme cases where $f(x) = \infty$ for *all* $x \in \mathbb{R}$ or $f(x) = -\infty$ for *some* $x \in \mathbb{R}$ can cause some inconveniences in what follows, so we exclude them. In the usual diction this means that we restrict ourselfs to *proper* convex functions.

Behind definition (A.1) a nice geometrical intuition is hidden:

> *To determine the value of $f^*(q)$ for some given q, take an infinite straight line with slope q and move it from below towards the function f until it touches the graph of f. The intersection of the line with the vertical axis then gives $-f^*(q)$.*

This principle is depicted in Figure A.1. Observe that it still works in extreme cases as, e.g., when for a given q there is no room below f where the straight line with slope q can be put. Then the line has to be moved infinitely far down and we get $f^*(q) = \infty$.

f^* itself is a convex function, but with the additional property of being closed, where for a *closed* convex function the *epigraph*

$$\mathrm{epi} f := \{(x, y) \in \mathbb{R}^2 \colon y \geq f(x)\},$$

is a closed set. Convex functions and epigraphs are in a one-to-one relation, so for any convex function f one can define its *closure* $\mathrm{cl}(f)$ by taking the convex function with epigraph $\mathrm{cl}(\mathrm{epi} f)$. It is the key property of convex conjugation that by applying the operation twice, one almost gets back to the original function in the sense that

$$f^{**} = \mathrm{cl}(f).$$

The convex conjugate f^* is also known as the *Legendre-* or *Fenchel-Legendre-transform* of f. This denotation is especially common in large deviations theory, where logarithmic moment generating functions and rate functions form convex conjugate pairs.

A.3 Some calculations

First we will show the following basic estimate for the tail of the standard-normal distribution.

Lemma 17. *For $t \geq 0$ we have*

$$\int_t^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \leq \frac{1}{2} \exp\left(-\frac{t^2}{2}\right).$$

Proof.

$$\begin{aligned}\int_t^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y+t)^2}{2}\right) dy \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2+2yt+t^2}{2}\right) dy \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \underbrace{\exp(-yt)}_{\leq 1} \exp\left(-\frac{t^2}{2}\right) dy \\ &\leq \exp\left(-\frac{t^2}{2}\right) \underbrace{\int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy}_{=\frac{1}{2}} = \frac{1}{2} \exp\left(-\frac{t^2}{2}\right).\end{aligned}$$

□

Another variant of this is

$$\int_t^\infty \exp\left(-\frac{u^2}{2\sigma^2}\right) du \leq \frac{\sqrt{2\pi\sigma^2}}{2} \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{A.2}$$

We show that for large enough c the integral in (4.19) converges to zero as $n \to \infty$. We have

$$\begin{aligned}&2(4n)^4 \int_{c(n\log n)^{1/2}}^\infty \exp\left(-\frac{(y-3c_2/2)^2}{128c_3^2 n}\right) dy \\ &\quad = 2(4n)^4 \int_{c(n\log n)^{1/2}-3c_2/2}^\infty \exp\left(-\frac{z^2}{128c_3^2 n}\right) dz \\ &\quad \leq (4n)^4 \sqrt{128\pi c_3^2 n} \exp\left(-\frac{(c(n\log n)^{1/2}-3c_2/2)^2}{128c_3^2 n}\right),\end{aligned}$$

where we used (A.2) with $\sigma^2 = 64c_3^2 n$ in the inequality. For n large enough certainly

$$(c(n\log n)^{1/2} - 3c_2/2)^2 \geq \frac{c^2 n \log n}{2},$$

so we get

$$\begin{aligned}&\leq (4n)^4 \sqrt{128\pi c_3^2 n} \exp\left(-\frac{(c(n\log n)^{1/2}-3c_2/2)^2}{128c_3^2 n}\right) \\ &\leq (4n)^4 \sqrt{128\pi c_3^2 n} \exp\left(-\frac{c^2 \log n}{256 c_3^2}\right) \\ &= (4n)^4 \sqrt{128\pi c_3^2 n} \cdot n^{-\frac{c^2}{256c_3^2}} \xrightarrow[n\to\infty]{} 0\end{aligned}$$

for c large enough.

A.4 A lemma for almost additive sequences

Lemma 18. *Let $(a_n)_{n\geq 0}$ be a sequence for which $\alpha := \lim_{n\to\infty} a_n/n$ exists and for which*

$$2a_n - l(2n) \leq a_{2n} \leq 2a_n + u(2n), \tag{A.3}$$

for all n and for some nonnegative functions $l(\cdot)$ and $u(\cdot)$. Then

$$-\sum_{i=1}^{\infty} 2^{-i} l(2^i n) \leq n\alpha - a_n \leq \sum_{i=1}^{\infty} 2^{-i} u(2^i n).$$

Proof. Iterate (A.3) to get for $k > 0$

$$2^k a_n - \sum_{j=0}^{k-1} 2^j l(2^{k-j} n) \leq a_{2^k n} \leq 2^k a_n + \sum_{j=0}^{k-1} 2^j u(2^{k-j} n).$$

Dividing by $2^k n$ and changing the index of summation to $i = k - j$ gives

$$\frac{a_n}{n} - \frac{1}{n}\sum_{i=1}^{k} 2^{-i} l(2^i n) \leq \frac{a_{2^k n}}{2^k n} \leq \frac{a_n}{n} + \frac{1}{n}\sum_{i=1}^{k} 2^{-i} u(2^i n).$$

The assertion follows by letting $k \to \infty$. □

A.5 More calculations

A.5.1 Proof of Lemma 16

We show that

$$\sum_{b\in B_{kn,m,kn\rho}} \binom{kn/m}{b_{-m},\ldots,b_m} \exp\left(\sum_{h=-m}^{m} b_h x_h\right) \leq$$
$$\leq C'' \left(\frac{kn}{m}\right)^{2m-\frac{1}{2}} e^{\frac{kn}{m}\tilde{f}_\rho(x)} \left\{1 + O\left(\left(\frac{kn}{m}\right)^{-1}\right)\right\}. \tag{A.4}$$

To start with we state an upper bound for multinomials which follows directly from Stirling's formula.

Lemma 19. *Let $k_1,\ldots,k_m, n \in \mathbb{N}$ with $\sum_{i=1}^m k_i = n$. Then, for $k_1,\ldots,k_m$ large enough we have*

$$\binom{n}{k_1,\ldots,k_m} \leq (\sqrt{2\pi})^{-(m-1)} \sqrt{n} e^{n\left\{-\sum_{i=1}^m \frac{k_i}{m}\log\frac{k_i}{m}\right\}} (1 + O(n^{-1})).$$

Proof. Recall that according to Stirling's formula

$$n! = \sqrt{2\pi} n^{n+1/2} e^{-n} \left(1 + \frac{1}{12n} + O(n^{-2})\right).$$

Plugging this into the definition of the multinomial and using

$$\left(\prod_{i=1}^{m} \left(1 + \frac{1}{12k_i} + O(k_i^{-2})\right)\right)^{-1} \leq 1$$

and $(\sqrt{k_1 \cdots k_m})^{-1} \leq 1$ gives the result. □

Inserting the result from Lemma 19 into the left-hand side of (A.4) and recalling the definition of the function f_x we get

$$\sum_{b \in B_{kn,m,kn\rho}} \binom{kn/m}{b_{-m}, \ldots, b_m} \exp\left(\sum_{h=-m}^{m} b_h x_h\right) \leq$$
$$\leq (\sqrt{2\pi})^{-2m} \sqrt{\frac{kn}{m}} \sum_{b \in B_{kn,m,kn\rho}} e^{\frac{kn}{m} f_x\left(\frac{m}{kn} \cdot b\right)} \{1 + O(m/(kn))\}. \quad \text{(A.5)}$$

The sum on the right hand side in (A.5) can be interpreted as an approximating Riemann sum to a $(2m-1)$-dimensional volume integral in the sense that

$$\left(\frac{kn}{m}\right)^{-(2m-1)} \sum_{b \in B_{kn,m,kn\rho}} e^{\frac{kn}{m} f_x\left(\frac{m}{kn} \cdot b\right)} \leq$$
$$\leq \int_{P_{m,\rho}} e^{\frac{kn}{m} f_x(\pi)} \mathrm{d}_{(2m-1)}(\pi) + E_n, \quad \text{(A.6)}$$

with some error term E_n. In this approximation the approximating cells have volume $(kn/m)^{-(2m-1)}$ and a diameter which is of the order of $(kn/m)^{-1}$. Since the integrand $e^{\frac{kn}{m} f_x(\pi)}$ grows exponentially in k, we can not deduce that the error term vanishes as $k \to \infty$. Anyway, observe that the maximum over $P_{m,\rho}$ of the total differential of the integrand can be bounded by

$$C(kn/m) e^{\frac{kn}{m} \tilde{f}_\rho(x)}$$

for some constant C. The error term can then be bounded by multiplying this upper bound by the diameter of the approximating cells which gives

$$E_n \leq C' e^{\frac{kn}{m} \tilde{f}_\rho(x)}$$

with some constant C'. Furthermore, a Laplace approximation of the integral gives that

$$\int_{P_{m,\rho}} e^{\frac{kn}{m} f_x(\pi)} \mathrm{d}_{(2m-1)}(\pi) = \\ = C'(\sqrt{2\pi})^{2m-1} \left(\frac{kn}{m}\right)^{-\frac{2m-1}{2}} e^{\frac{kn}{m}\tilde{f}_\rho(x)}\{1 + O(m/(kn))\}.$$

Collecting these bounds together now shows that

$$\sum_{b \in B_{kn,m,kn\rho}} e^{\frac{kn}{m} f_x(\frac{m}{kn} \cdot b)} \leq C \left(\frac{kn}{m}\right)^{2m-1} e^{\frac{kn}{m}\tilde{f}_\rho(x)} \left\{1 + O\left(\left(\frac{kn}{m}\right)^{-\frac{2m-1}{2}}\right)\right\}.$$

Observe that the leading term in this upper bound merely came from the error term E_n rather than from the Laplace approximation of the integral. The proof of (A.4) now follows directly from (A.5).

A.5.2 Calculation of $\tilde{f}_\rho(x)$

Recall the definition of the function $f_x \colon \mathbb{R}^{2m+1} \to \mathbb{R}$ as

$$f_x(\pi_{-m}, \ldots, \pi_m) := -\sum_{h=-m}^{m} \pi_h[\log(\pi_h) - x_h],$$

where $x = (x_{-m}, \ldots, x_m) \in \mathbb{R}^{2m+1}$. For $\rho \in [-1, 1]$ we now calculate

$$\tilde{f}_\rho(x) := \max\{f_x(\pi) \colon \pi \in P_{m,\rho}\},$$

where

$$P_{m,\rho} := \left\{\pi = (\pi_{-m}, \ldots, \pi_m) \in [0,1]^{2m+1} \colon \sum_{h=-m}^{m} \pi_h = 1, \sum_{h=-m}^{m} h\pi_h = \rho m\right\}.$$

From the method of Lagrange multipliers we see that for all $-m \leq h \leq m$ we have to equate

$$\frac{\partial}{\partial \pi_h}\left\{f_x(\pi) + \eta_1\left[\sum_{h=-m}^{m} \pi_h - 1\right] + \eta_2\left[\sum_{h=-m}^{m} h\pi_h - \rho m\right]\right\} = \\ = \log \pi_h - x_h + 1 + \eta_1 + \eta_2 h$$

to zero which gives that π is of the form

$$\pi_h = C_{x,\eta_2}^{-1} e^{x_h - \eta_2 h} \tag{A.7}$$

with the normalizing constant

$$C_{x,\eta_2} := \sum_{h=-m}^{m} e^{x_h - \eta_2 h}.$$

Let us first examine the special case of $\eta_2 = 0$. In this case $\hat{\pi} := (\hat{\pi}_{-m}, \ldots, \hat{\pi}_m)$, where

$$\hat{\pi}_h := C_{x,0}^{-1} e^{x_h}$$

maximizes $f_x(\pi)$ in $P_{m,\hat{\rho}}$, where

$$\hat{\rho} := \frac{1}{m} \sum_{h=-m}^{m} h\hat{\pi}_h.$$

Especially for symmetric x, also $\hat{\pi}$ is symmetric and it follows that $\hat{\rho} = 0$.

To get the solution for general ρ, consider the distribution $\hat{\mu}$ which is defined by $\hat{\pi}$ on the set $\{-m, \ldots, m\}$. As can be seen from the basic form (A.7), the solution arises from an exponential tilting of $\hat{\mu}$. To work out the details, define the logarithmic moment generating function $\hat{\Lambda}$ of $\hat{\mu}$ as

$$\hat{\Lambda}(\lambda) := \log \left(\sum_{h=-m}^{m} \hat{\pi}_h e^{\lambda h} \right).$$

For the given ρ determine the solution λ_ρ of

$$\hat{\Lambda}'(\lambda) = \rho m$$

and define

$$\bar{\pi}_h := \hat{\pi}_h e^{\lambda_\rho h - \hat{\Lambda}(\lambda_\rho)}.$$

Then $\bar{\pi}$ maximizes $f_x(\pi)$ in $P_{m,\rho}$.

All that remains is to calculate $\tilde{f}_\rho(x) = f_x(\bar{\pi})$. To do this, observe that a short calculation shows that

$$\log \bar{\pi}_h = x_h + \lambda_\rho h - \log \left(\sum_{h'=-m}^{m} e^{x_{h'} + \lambda_\rho h'} \right),$$

which gives

$$\tilde{f}_\rho(x) = \log \left(\sum_{h'=-m}^{m} e^{x_{h'} + \lambda_\rho h'} \right) - m\rho\lambda_\rho.$$

Bibliography

[AG96] S. F. Altschul and W. Gish, *Local alignment statistics*, Methods in Enzymology **266** (1996), 460–480.

[AGlG89] R. Arratia, L. Goldstein, and l. Gordon, *Two moments suffice for Poisson approximations: The Chen-Stein method*, Ann. Prob. **17** (1989), no. 1, 9–25.

[AGM+90] S. F. Altschul, W. Gish, W. Miller, E.W. Myers, and D. J. Lipman, *Basic local alignment search tool*, J. Mol. Biol. **215** (1990), no. 3, 403–410.

[Ale93] K. S. Alexander, *A note on some rates of convergence in first-passage percolation*, Ann. Appl. Prob. **3** (1993), no. 1, 81–90.

[Ale94] K. S. Alexander, *The rate of convergence of the mean length of the longest common subsequence*, Ann. Appl. Prob. **4** (1994), no. 4, 1074–1082.

[Ale97] K. S. Alexander, *Approximation of subadditive functions and convergence rates in limiting-shape results*, Ann. Prob. **25** (1997), no. 1, 30–55.

[AW94] R. Arratia and M. S. Waterman, *A phase transition for the score in matching random sequences allowing deletions*, Ann. Appl. Prob. **4** (1994), no. 1, 200–225.

[Azu67] K. Azuma, *Weighted sums of certain dependent random variables*, Tôhoku Math. J. **19** (1967), 357–367.

[BR60] R. R. Bahadur and R. Ranga Rao, *On deviations of the sample mean*, Ann. Math. Stat. **31** (1960), 1015–1027.

[Bun02] R. Bundschuh, *Rapid significance estimation in local sequence alignment with gaps*, J. Comp. Biol. **9** (2002), 243–260.

[CLR90] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest, *Introduction to algorithms*, The MIT Electrical Engineering and Computer Science Series, MIT Press, Cambridge, MA, 1990.

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

[dH00] F. den Hollander, *Large deviations*, Fields Institute Monographs, vol. 14, American Mathematical Society, Providence, RI, 2000.

[DKZ94a] A. Dembo, S. Karlin, and O. Zeitouni, *Critical phenomena for sequence matching with scoring*, Ann. Prob. **22** (1994), no. 4, 1993–2021.

[DKZ94b] A. Dembo, S. Karlin, and O. Zeitouni, *Limit distributions of maximal non-aligned two-sequence segmental score*, Ann. Prob. **22** (1994), no. 4, 2022–2039.

[DSO78] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, *A model of evolutionary change in proteins*, Atlas of protein sequence and structure (M. O. Dayhoff and R. V. Eck, eds.), vol. 5, Natl. Biomed. Res. Found., 1978, pp. 345–352.

[Dur96] R. Durrett, *Probability: Theory and examples*, 2nd ed., Duxbury Press, Belmont, 1996.

[DZ93] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Jones and Bartlett, 1993.

[EKM97] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling extremal events*, Springer, 1997.

[Gri84] G. R. Grimmett, *Large deviations in subadditive processes and first-passage percolation*, Particle systems, random media and large deviations (R. Durrett, ed.), 1984.

[Ham74] J. M. Hammersley, *Postulates for subadditive processes*, Ann. Prob. **2** (1974), no. 4, 652–680.

[HH92] S. Henikoff and J. G. Henikoff, *Amino acid substitution matrices from protein blocks*, Proc. Natl. Acad. Sci. USA **89** (1992), 10915–10919.

[Hir98] K. Hirano, *Determination of the limiting coefficient for exponential functionals of random walks with positive drift*, J. Math. Sci. Univ. Tokyo **5** (1998), 299–332.

[Hoe63] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc. **58** (1963), 13–30.

[HW65] J. M. Hammersley and J. A. D. Welsh, *First-passage percolation, subadditive processes, stochastic networks and generalized renewal theory*, Bernoulli-Bayes-Laplace Anniversary Volume (J. Neyman and L. M. Le Cam, eds.), Springer-Verlag, Berlin, 1965.

[Igl72] D. L. Iglehart, *Extreme values in the GI/G/1 queue*, Ann. Math. Stat. **43** (1972), 627–635.

[JC69] T. H. Jukes and C. R. Cantor, *Evolution of protein molecules*, Mammalian Protein Metabolism (H. N. Munro, ed.), Academic Press, 1969, pp. 21–132.

[KD92] S. Karlin and A. Dembo, *Limit distributions of maximal segmental score among Markov-dependent partial sums*, Adv. Appl. Prob. **24** (1992), 113–140.

[Kes84] H. Kesten, *Aspects of first-passage percolation*, École d'Été de probabilités de Saint-Flour XIV (Berlin, Heidelberg, New York, Tokyo) (P. L. Hennequin, ed.), Lecture Notes in Mathematics, vol. 1180, Springer-Verlag, 1984, pp. 125–264.

[Kin75] J. F. C. Kingman, *Subadditive processes*, École d'Été de probabilités de Saint-Flour V (Berlin, Heidelberg, New York, Tokyo) (P. L. Hennequin, ed.), Lecture Notes in Mathematics, vol. 539, Springer-Verlag, 1975, pp. 167–223.

[KS91] J. Krug and H. Spohn, *Kinetic roughening of growing surfaces*, Solids far from equilibrium: Growth, morphology and defects (C. Godrèche, ed.), Cambridge University Press, 1991, pp. 479–582.

[MGW02] D. Metzler, S. Grossmann, and A. Wakolbinger, *A Poisson point model featuring gapped local alignments*, Stat. Prob. Lett. **60** (2002), no. 1, 91–100.

[Mot02] R. Mott, *Accurate formula for p-values of gapped local sequence and profile alignments*, J. Mol. Biol. **300** (2002), 649–659.

[NP95] C. M. Newman and M. S. T. Piza, *Divergence of shape fluctuations in two dimensions*, Ann. Prob. **23** (1995), no. 3, 977–1005.

[NW70] S. B. Needleman and C. D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequences of two proteins*, J. Mol. Biol. **48** (1970), 444–453.

[OBH99] R. Olsen, R. Bundschuh, and T. Hwa, *Rapid assessment of extremal statistics for local alignment with gaps*, Proceedings of

The Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB99) (T. Lengauer et al., ed.), AAAI Press, Menlo Park, 1999, pp. 211–222.

[Pea98] W. R. Pearson, *Empirical statistical estimates for sequence similarity searches*, J. Mol. Biol. **276** (1998), 71–84.

[PL88] W. R. Pearson and D. J. Lipman, *Improved tools for biological sequence comparison*, Proc. Natl. Acad. Sci. USA **85** (1988), no. 8, 2444–2448.

[Rhe95] W. T. Rhee, *On rates of convergence for common subsequences and first passage time*, Ann. Appl. Prob. **5** (1995), no. 1, 44–48.

[Roc79] R. T. Rockafellar, *Convex analysis*, Princeton University Press, 1979.

[RR91] A. B. Robinson and L. R. Robinson, *Distribution of glutamine and asparagine residues and their near neighbors in peptids and proteines*, Proc. Natl. Acad. Sci. USA **88** (1991), no. 20, 8880–8884.

[SS01] J. Storey and D. Siegmund, *Approximate p-values for local sequence alignments: Numerical studies*, J. Comp. Biol. **8** (2001), no. 5, 549–556.

[Ste97] J. M. Steele, *Probability theory and combinatorial optimization*, SIAM, 1997.

[SV98] R. Spang and M. Vingron, *Statistics of large-scale sequence searching*, Bioinformatics **14** (1998), 279–284.

[SW78] R. T. Smythe and J. C. Wierman, *First-passage percolation on the square lattice*, Springer, 1978.

[SW81] T. F. Smith and M. S. Waterman, *Identification of common molecular subsequences*, J. Mol. Biol. **147** (1981), 195–197.

[SY00] D. Siegmund and B. Yakir, *Approximate p-values for local sequence alignments*, Ann. Stat. **28** (2000), no. 3, 657–680.

[Tal96] M. Talagrand, *A new look at independence*, Ann. Prob. **24** (1996), no. 1, 1–34.

[TKF91] J. L. Thorne, H. Kishino, and J. Felsenstein, *An evolutionary model for maximum likelihood alignment of DNA*, J. Mol. Evol. **33** (1991), 114–124.

[WV94a] M. S. Waterman and M. Vingron, *Rapid and accurate estimates of statistical significance for sequence database searches*, Proc. Natl. Acad. Sci. USA **91** (1994), no. 11, 4625–4628.

[WV94b] M. S. Waterman and M. Vingron, *Sequence comparison significance and Poisson approximation*, Stat. Science **9** (1994), no. 3, 367–381.

[Zha95] Y. Zhang, *A limit theorem for matching random sequences allowing deletions*, Ann. Appl. Prob. **5** (1995), no. 4, 1236–1240.

Lebenslauf

Steffen Grossmann
geboren am 21. Januar 1969 in Offenbach am Main,

1975 bis 1988	Schulausbildung in Mühlheim am Main
1988	Erwerb der Allgemeinen Hochschulreife am Friedrich Ebert-Gymnasium, Mühlheim am Main Leistungskurse: Mathematik und Französisch
August 1988 bis März 1990	Zivildienst in Mühlheim in der Individuellen Schwerstbehinderten-Betreuung
Oktober 1990 bis November 1997	Studium der Mathematik mit Nebenfach Philosophie an der Johann Wolfgang Goethe-Universität in Frankfurt am Main
1990 bis 1992	Grundstudium bei den Professoren H. Behr, W. K. Essler, K. H. Müller, A. Wakolbinger und J. Weidmann
Oktober 1992	Diplom-Vorprüfung
1992 bis 1997	Hauptstudium bei den Professoren H. Behr, J. Bliedtner, F. Kambartel, G. Kersting und A. Wakolbinger Diplomarbeit bei Prof. A. Wakolbinger zum Thema *Äquipolarität von Galton Watson-Bäumen und 2. Momente* (Zweitgutachter: Prof. H. Dinges)
November 1997	Diplom-Hauptprüfung
Seit 1. April 1998	wissenschaftlicher Mitarbeiter des *Instituts für Stochastik und mathematische Informatik* am FB Mathematik der Universität Frankfurt (bei Prof. H. Dinges) Bearbeitung des Promotionsthemas bei Prof. A. Wakolbinger