University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Rasmus Erlemann

Effects of Parameters Choice in Random Sequence Comparison

Mathematics and Statistics Curriculum
(Mathematics)
Master's Thesis (30 ECTS)

Supervisor: Jüri Lember

Tartu 2017

Effects of Parameters Choice in Random Sequence Comparison

Master's Thesis
Rasmus Erlemann

Abstract. The aim of this thesis is to generalize the results from the article [LMT] for the score of mismatched letters. This is an introductory work and it includes complete theoretical build-up of random sequence comparison, depending on parameters. At the end we prove the large deviations inequality for the proportions of mismatches. It gives probability for proportions of mismatches being near the asymptotic proportions of mismatches. This thesis also includes a fair bit of practical examples and an introduction into algorithms, used for sequence comparison.

Keywords. Sequences, simulation, ergodic transformations, graphs.
CERCS code: P160 Statistics, operation research, programming, actuarial mathematics.

Parameetrite valiku mõju juhuslike jadade võrdlemises

Magistritöö
Rasmus Erlemann

Lühikokkuvõte. Käesoleva magistritöö eesmärk on artikli [LMT] tulemuste üldistamine joondamata tähtede skoori jaoks. Töö on sissejuhatav ja sisaldab terviklikku teooria ülesehitust parameetritest sõltuva juhuslike jadade võrdlemiseks. Viimases peatükis tuleb juttu suurte hälvete printsiibist. See annab tõkke tõenäosusele, et joondamata tähtede proportsioonid on asümptootiliste joondamata proportsioonide lähedal. Töö sisaldab ka mitmeid näiteid ning sissejuhatust algoritmidesse, mida kasutatakse jadade võrdlemises.

Märksõnad. Jadad, simulatsioon, ergoodilised teisendused, graafid.
CERCS kood: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Contents

Introduction

Finding the best alignment of two DNA, RNA or amino acid sequences has become almost the standard technique for sequence comparison in molecular biology. It is used to determine whether and where two sequences are similar (homologous), to determine evolutionary history between species, to find consensus sequences, and other significant functions. However, in present methods for evaluating optimal sequence alignment, specific score function (which determines the relationship between sequence elements) and gap penalty must be specified. Therefore, the (biological) significance of the optimal alignment depends heavily upon the "right" choice of parameters. There is considerable disagreement among molecular biologists about the correct choice and it is probably the case that there is no unique choice for the parameters.

The significance of an alignment is based either on biological grounds, or on its sensitivity to the choice of parameters. Instead of repeatedly varying the parameter weights, we need to restrict the domain of parameters by using estimation of their sensitivity. As an example (see, e.g., [SK], pages $290 - 293$) to demonstrate the difficulties in finding the relative parameter weights by a specific example: the comparison of human and E. Coli 5S RNA (two sequences of 120 characters each over a four-letter alphabet). The solution of [SK] involves varying one parameter, the number of indels, until its approriate value is found. Similarly [FS] demonstrates how a biologically accepted alignment may easily be missed if inappropriate weights are used. The main aim of this thesis is to prove concentration inequalities for the sensitivity of varying parameters. This can be used to restrict the domain of parameters that are taken into consideration.

For simplicity, we mainly restrict our theory by only considering binary alphabets. This restriction can most likely be partially eliminated, as [BSS] proved that binary and non-binary alphabets enjoy many similar properties.

This thesis is composed of 5 chapters.

The first chapter introduces basic theory of random sequence comparison and other preliminary ideas. It also consists of several examples to illustrate the fundamental theory. The second part of the first chapter is devoted to introducing alignment graphs. There, we deduce the apparatus to do sequence comparison in

practice.

In the second chapter we investigate the effect of varying gap price in sequence comparison.

In the third chapter, we generalize the theory that has been introduced in the second chapter. We derive similar results for the effect of varying score of mismatch, instead of gap price. This allows us the investigate the effect of varying both variables (mismatch score and gap price).

In the fourth chapter, we explore the asymptotic properties of random sequences, as the length of the sequences goes to infinity. We derive asymptotic results for the proportions (mismatches and gaps) and latter part of the chapter is dedicated to confirming the results by running simulations.

The fifth chapter is devoted to deriving large deviations inequality for proportions. It allows us to estimate how varying mismatch proportions affects the optimal alignment score.

In the sixth chapter we prove that the two-variable function B_n can be reduced to the one-variable case. This allows us to prove some basic properties of the two-variable B_n, while varying only one variable.

When constructing examples, we have used listings from appendix A. Some examples required slight alteration in code, but changes were trivial and interested reader can easily reconstruct them.

This thesis is based on author's own work. Still, it heavily relies on [LMT] and the supervisor has strongly contributed to proving the results and laying out the overall backbone of the work.

Chapter 1

Preliminaries

In this chapter we recall supplementary notation and results needed in most of the following chapter. These include basics of sequence comparison and introduction to alignment graphs. This chapter relies mostly on [LMT].

1.1 Basics of Sequence Comparison

Throughout this thesis $X = (X_1, X_2, \ldots, X_n)$, $Y = (Y_1, Y_2, \ldots, Y_n)$ and $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$, with $n \in \mathbb{N}$ being the length of the sequences, are either two fixed sequences or pairwise independent sequences of iid (independent identically distributed) random variables. We reserve x, y for fixed sequences and X, Y for sequences of random variables. We also let $X_1 \sim Y_1$. Our goal is to develop a model to quantify the difference between those two sequences.

Each element of the sequences X, Y and x, y is drawn from a finite set $\mathbb{A}$, which we call an alphabet. Elements of $\mathbb{A}$ are called letters.

Definition 1.1. Symmetric and non-constant functional $S\colon \mathbb{A} \times \mathbb{A} \to \mathbb{R}$ is called a **pairwise scoring function**.

Pairwise scoring function's role is to quantify the relationship between sequences x and y elements. Next example illustrates it.

Example 1.2. Let us fix the alphabet $\mathbb{A} = \{a, b, c\}$, sequences $x = (a, a, b, b, c)$, $y = (b, b, a, a, c)$ and the pairwise scoring function as

$$\begin{aligned} &S(a,a) = 1, && S(a,b) = 0, && S(a,c) = 0,\\ &S(b,a) = 0, && S(b,b) = 1, && S(b,c) = 0,\\ &S(c,a) = 0, && S(c,b) = 0, && S(c,c) = 1. \end{aligned}$$

Pairwise scores of x and y in-line elements are

x	a	a	b	b	c
y	b	b	a	a	c
$S(\cdot,\cdot)$	0	0	0	0	1

We denote

$$F := \max_{(a,b)\in\mathbb{A}\times\mathbb{A}} S(a,b),$$

as the largest possible score.

Let $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ be two strictly increasing sequences of natural numbers, i.e. $1 \leq \pi_1 < \pi_2 < \ldots < \pi_k \leq n$ and $0 \leq k \leq n$.

Definition 1.3. Subsequences $(X_{\pi_1}, X_{\pi_2}, \ldots, X_{\pi_k})$ and $(Y_{\mu_1}, Y_{\mu_2}, \ldots, Y_{\mu_k})$ or $(x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_k})$ and $(y_{\mu_1}, y_{\mu_2}, \ldots, y_{\mu_k})$ are called the **sequences of aligned letters**. Remaining letters are called **misaligned**.

It should be noted that k is the number of aligned letters. An alignment consists of aligned and misaligned letters. We represent it in a tabular form and set each misaligned letter in correspondence to an indel (we denote it by $-$). This allows us to leave some elements of the sequences "out of comparison" and we are no longer obligated to only compare in-line elements.

Example 1.4. Let us fix the alphabet $\mathbb{A} = \{a, b\}$ and sequences

$$x = (a, b),$$
$$y = (b, b).$$

All possible alignments are

x	a	b
y	b	b

,

x	a	$-$	b
y	b	b	$-$

,

x	a	b	$-$
y	$-$	b	b

,

x	$-$	a	b
y	b	b	$-$

,

x	a	$-$	b
y	$-$	b	b

,

x	a	b	$-$	$-$
y	$-$	$-$	b	b

.

We determine each alignment by sequences π and μ. For example we do not differentiate between the following alignments:

x	a	b	$-$
y	b	$-$	b

,

x	a	$-$	b
y	b	b	$-$

.

We denote the set of all possible alignments, with $n \in \mathbb{N}$ being the length of sequences, by

$$\overline{\mathcal{O}}_n := \{(\pi, \mu) \in : 1 \leq \pi_1 < \pi_2 < \ldots \pi_k \leq n,\ 1 \leq \mu_1 < \mu_2 < \ldots < \mu_k \leq n, 0 \leq k \leq n\}.$$

Definition 1.5. The number

$$q(\pi, \mu) := \frac{n-k}{n} \in [0, 1]$$

is the **proportion of gaps** of the alignment (π, μ).

The average score of aligned letters is defined by

$$t(\pi, \mu) := \frac{1}{k} \sum_{i=1}^{k} S(x_{\pi_i}, y_{\mu_i}).$$

Note that our definition of gap slightly differs from the one that is commonly used in the sequence alignment literature, where a gap consists of maximal number of consecutive indels (insertion and deletion) in one side. Our gap actually corresponds to a pair of indels, one in x-side and another in y-side. Since we consider the sequences of equal length, to every indel in x-side corresponds an indel in y-side, so considering them pairwise is justified. In other words, the number of gaps in our sense is the number of indels in one sequence. We also consider a gap price δ.

Definition 1.6. Given a pairwise scoring function S and the gap price δ, **score of the alignment** (π, μ), when aligning x and y is defined by

$$U^{\delta}_{(\pi,\mu)}(x, y) := \sum_{i=1}^{k} S(x_{\pi_i}, y_{\mu_i}) + \delta(n - k).$$

Score of the alignment can be written down as the convex combination

$$U^{\delta}_{(\pi,\mu)}(x, y) = n(t(\pi, \mu)(1 - q(\pi, \mu)) + \delta q(\pi, \mu)). \tag{1.1}$$

In our general scoring scheme δ can also be positive, although usually $\delta \leq 0$, penalizing the mismatch. For negative δ, the quantity $-\delta$ is usually called the gap penalty.

Definition 1.7. The **optimal alignment score** of x and y is defined to be

$$L_n(\delta) := \max_{(\pi,\mu)\in\overline{\mathcal{O}}_n} U^{\delta}_{(\pi,\mu)}(x, y).$$

The alignments achieving the maximum are called optimal. For every $\delta \in \mathbb{R}$, let us denote

$$B_n(\delta) := \frac{L_n(\delta)}{n}.$$

Note that to every alignment (π, μ) corresponds an unique pair $(t(\pi, \mu), q(\pi, \mu))$, but different alignments can have the same $t(\pi, \mu)$ and $q(\pi, \mu)$, thus from (1.1) we get that

$$B_n(\delta) = \max_{(\pi,\mu)\in\overline{\mathcal{O}}_n} (t(\pi, \mu)(1 - q(\pi, \mu)) + \delta q(\pi, \mu)) = \max_{(t,q)\in\overline{\mathcal{O}}'_n} (t(1 - q) + \delta q), \tag{1.2}$$

where $\overline{\mathcal{O}}'_n$ consists of all possible possible pairs (t, q). In other words, while $\overline{\mathcal{O}}_n$ consists of all possible subsequences that produce alignments, $\overline{\mathcal{O}}'_n$ consists of all possible gap proportions and average scores of aligned letters. We usually leave out arguments (π, μ) and use (t, q) to refer to an alignment.

Example 1.8. Let us fix sequences

$$x = (a, b, a, a, b),$$
$$y = (b, a, a, b, b)$$

and the score function as $S(a, a) = 1,\ S(b, b) = 1,\ S(a, b) = 0$. Next, we will sketch the graph of the function B_5.

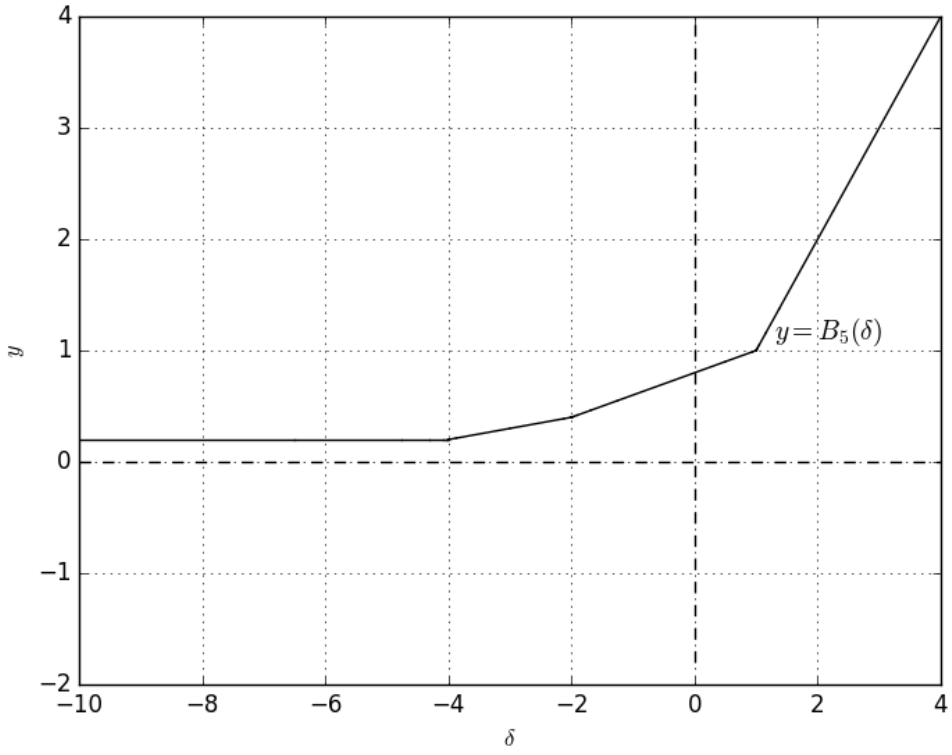

Figure 1.9: Graph of the function B_5

If we fix $\delta = -1$, then $B_5(\delta) = \dfrac{3}{5}$ and all optimal alignments are

x	a	b	a	a	$-$	b
y	$-$	b	a	a	b	b

, and

x	a	b	a	a	b	$-$
y	$-$	b	a	a	b	b

.

In what follows, we identify alignments with pairs (t, q), such that a pair (t, q) always corresponds to an alignment (π, μ) of x and y. Let us denote $\mathcal{O}_n(\delta)$ for each $\delta \in \mathbb{R}$ as the set of optimal pairs, i.e. a pair $(t, q) \in \mathcal{O}_n(\delta)$ if and only if $t(1 - q) + \delta q = B_n(\delta)$ and $(t, q) \in \overline{\mathcal{O}}'_n$. Note that the set $\mathcal{O}_n(\delta)$ is not necessarily a singleton. Let us denote

$$\underline{q}_n(\delta) := \min\{q \colon (t, q) \in \mathcal{O}_n(\delta)\},$$

$$\overline{q}_n(\delta) := \max\{q\colon (t,q) \in \mathcal{O}_n(\delta)\}.$$

If $\underline{q}_n(\delta) = \overline{q}_n(\delta)$, we denote it by

$$q_n(\delta) := \underline{q}_n(\delta) = \overline{q}_n(\delta).$$

Example 1.10. Continuing Example 1.8, we can sketch the graph of q_5.

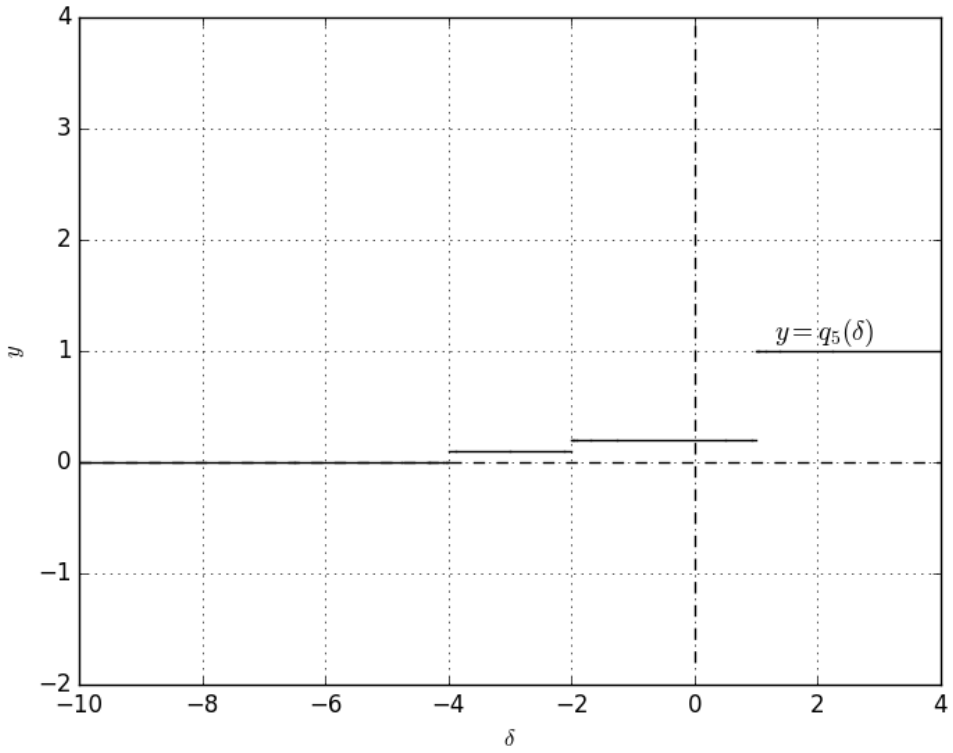

Figure 1.11: Graph of the function q_5

If we take $\delta = 1$, then $\underline{q}_5(\delta) \neq \overline{q}_5(\delta)$. All optimal alignments are

x	a	b	a	a	b	$-$
y	$-$	b	a	a	b	b

x	$-$	$-$	$-$	$-$	a	b	a	a	b
y	b	a	a	b	$-$	$-$	$-$	$-$	b

x	$-$	$-$	a	b	a	a	b
y	b	a	a	b	$-$	$-$	b

x	$-$	$-$	$-$	$-$	a	b	a	a	b
y	b	a	a	b	$-$	b	$-$	$-$	$-$

x	$-$	$-$	$-$	a	b	a	a	b
y	b	a	a	$-$	b	$-$	$-$	b

x	$-$	$-$	$-$	$-$	$-$	a	b	a	a	b
y	b	a	a	b	b	$-$	$-$	$-$	$-$	$-$

Since each optimal alignment has either $1, 2, 3, 4$ or 5 gaps, we can conclude that $\underline{q}_5(\delta) = \dfrac{1}{5}$ and $\overline{q}_5(\delta) = 1$. That is in accordance with Figure 1.11.

1.2 Alignment Graphs

Finding optimal alignments for longer sequences by hand is practically unreasonable. In this thesis, we use dynamic programming algorithm called Needleman-Wunsch algorithm to compute optimal alignments of two sequences

$$x = (x_1, x_2, \ldots, x_n),$$
$$y = (y_1, y_2, \ldots, y_n).$$

It can be viewed as a procedure for finding a maximum-weight path in a weighted alignment graph G. The nodes of G are arranged in an $(n+1) \times (n+1)$ grid (length of both sequences are n); rows (columns) are numbered consecutively from top to bottom (left to right) from 0 to n. We denote the nodes of G by their coordinates (i, j). Every node has an edge directed to its right neighbor and an edge to its neighbor below it. These edges have weight δ and they represent an indel in the alignment. Additionally, for $i, j \in \{1, 2, \ldots, n\}$, there is a diagonal edge directed into vertex (i, j) from vertex $(i-1, j-1)$. The weight of the edge is $S(x_i, y_j)$ and it represents an aligned letter pair in the alignment.

Each path Γ from $(0, 0)$ to (n, n) corresponds to a unique alignment. Hence, we will only consider paths with moves between node (i, j) to $(i, j+1)$, $(i+1, j)$ or $(i+1, j+1)$, for $i, j \in \{0, 1, 2, \ldots, n-1\}$. All possible paths are displayed in the Figure 1.12.

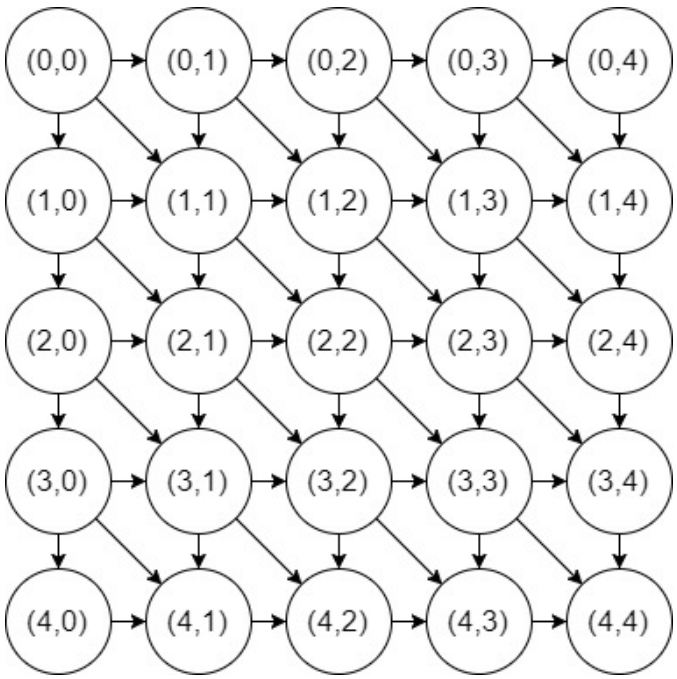

Figure 1.12: Paths in an alignment graph

Figure 1.13 displays the alignment graph for $x = (b, b, a, c, b, a, c, a, a, b)$ and $y = (a, b, b, c, a, b, b, a, a, c)$. Shaded areas are match blocks, which means that there is

a matching letter pair in the corresponding position in the alignment. The path Γ shown, corresponds to the alignment

x	$-$	b	b	a	c	$-$	b	$-$	a	c	a	a	b
y	a	b	b	$-$	c	a	b	b	a	$-$	a	$-$	c

.

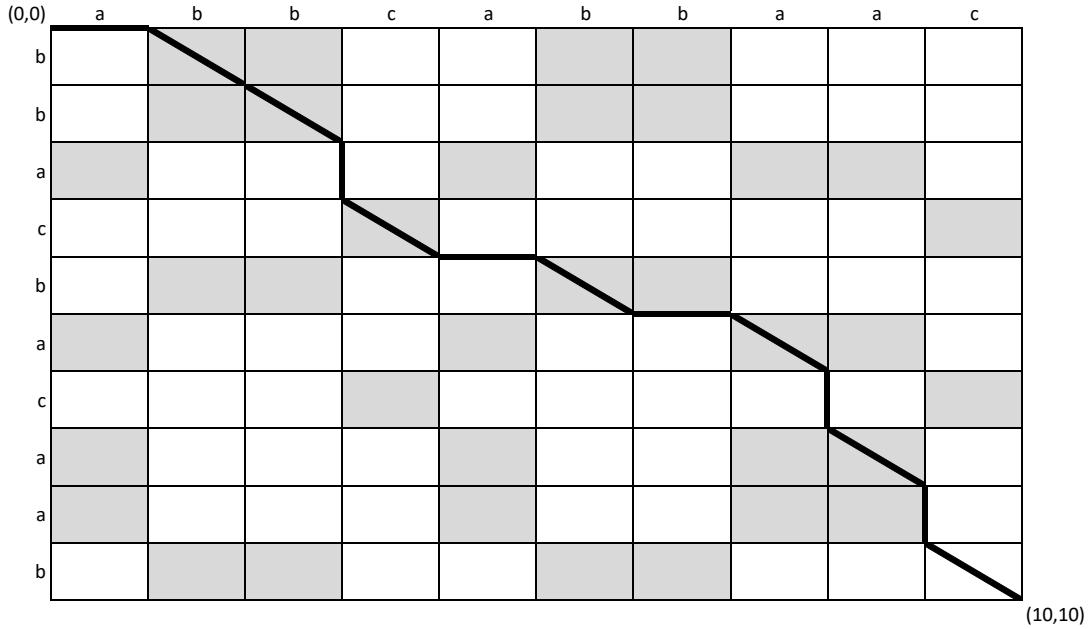

Figure 1.13: The alignment graph

Figure 1.13 can be made more thorough using different colors. For example, mismatch of letters a, b and b, c can be of different colors. This way, the model can be generalized for all different score functions.

Following pseudocode describes how Needleman-Wunsch algorithm implements alignment graphs to find the optimal alignment score. It also prints an optimal alignment as a side effect.

Algorithm 1: Neeldeman-Wunsch

Input: Two sequences x and y with length n
Output: Optimal alignment score α

for $i = 0, 1, 2, \ldots, n$ **do**
  $F(i, 0) \leftarrow i\delta$
end for
for $j = 0, 1, 2, \ldots, n$ **do**
  $F(0, j) \leftarrow j\delta$
end for
for $i = 1, 2, \ldots, n$ **do**
  for $j = 1, 2, \ldots, n$ **do**
    $F(i, j) \leftarrow \max\{F(i-1, j-1) + S(x_i, y_j), F(i-1, j) - \delta, F(i, j-1) - \delta\}$
    Set backtrace $T(i, j)$ to the maximizing pair (i', j')
  end for
end for
The score is $\alpha \leftarrow F(n, n)$
repeat
  if $T(i, j) = (i-1, j-1)$ **then**
    print $\begin{pmatrix} x_i \\ y_j \end{pmatrix}$
  else if $T(i, j) = (i-1, j)$ **then**
    print $\begin{pmatrix} x_i \\ - \end{pmatrix}$
  else
    print $\begin{pmatrix} - \\ y_j \end{pmatrix}$
  end if
  Set $(i, j) \leftarrow T(i, j)$
until $(i, j) = (0, 0)$.

Algorithm 1 stores $(n + 1) \times (n + 1)$ numbers. Each number takes a constant number of calculations to compute: three sums and a max. Hence, for filling the matrix F, the algorithm requires $O(n^2)$ time and memory. Given the filled matrix, the construction of the alignment is done in time $O(n)$. In this thesis, we restricted use of algorithm 1 to $n = 1000$.

In order to sketch the graph of the function B_n (depending on one variable), the author developed an algorithm, that sketches the graph with an adjustable error bound. Error bound describes the maximum difference between the approximated and the exact graph at each interval's middle point.

Algorithm 2: Graph B_n

Input: Interval $[a,b]$, $B_n\Big|_a$, $B_n\Big|_b$ and error bound ε
Output: Graph of B_n
Calculate $B_n\Big|_{\frac{a+b}{2}}$
if $\left|\dfrac{B_n\Big|_a - B_n\Big|_b}{2} - B_n\Big|_{\frac{a+b}{2}}\right| > \varepsilon$ **then**
Call algorithm 2 with Interval $\left[a, \dfrac{a+b}{2}\right]$, $B_n\Big|_a$, $B_n\Big|_{\frac{a+b}{2}}$ and error bound ε
Call algorithm 2 with Interval $\left[\dfrac{a+b}{2}, b\right]$, $B_n\Big|_{\frac{a+b}{2}}$, $B_n\Big|_b$ and error bound ε
else
Sketch a straight line through $B_n\Big|_a$ and $B_n\Big|_b$ on $[a,b]$
end if

1.3 Miscellaneous

Definition 1.14. Let $X_1, X_2, \ldots$ be a sequence of random variables defined on a probability space Ω. We say $X_1, X_2, \ldots$ **almost surely converges** (denoted a.s.) to a random variable X, defined on Ω, if

$$\mathbf{P}\left(\left\{\omega \in \Omega \colon \lim_n X_n(\omega) \neq X(\omega)\right\}\right) = 0.$$

In this thesis, when considering properties on a plane, we say that property A almost everywhere holds (a.e.), if

$$\mathrm{Leb}(\{x \in \mathbb{R}^2 \colon A(x) \text{ does not hold}\}) = 0.$$

Definition 1.15. Function $f \colon D \to \mathbb{R}$ is called **affine**, if there exists a linear function $g \colon D \to \mathbb{R}$ and a constant $C \in \mathbb{R}$, such that for all $x \in D$

$$f(x) = g(x) + C.$$

Chapter 2

Gap price as a variable

In this chapter we introduce basic tools for analysing the change of optimal alignments, when varying the gap price. Our main focus is on the score of the optimal alignment and its rate of change in relation to the gap price. Examples in this chapter help with visualizing and confirming the theory.

Lemma 2.1 (see, e.g., [LMT] claim 2.1)**.** *The function $\delta \mapsto B_n(\delta)$ is non-decreasing, piecewise linear and convex.*

Apart from B_n being piecewise linear, it should be mentioned that the number of intervals, in which B_n is linear, is finite. This is a result of sequences and alphabet being finite.

Lemma 2.2 (see, e.g., [CZ], pages 6-7)**.** *Let sequences x and y be of length n. The number of different alignments*[1] *is*

$$\sum_{k=n}^{2n} \binom{k}{n} \binom{n}{2n-k}.$$

Example 2.3. The following table demonstrates the number of alignments, depending on the size of the sequences.

n	1	2	3	4	5	6	7	8
number of alignments	3	13	63	321	1683	8989	48639	265729

Table 2.4: Number of alignments

It should be noted that calculating $B_n(\delta)$ without an efficient algorithm is practically impossible. Even for small sequence sizes ($n \approx 200$), the number of alignments is greater than the number of atoms in the universe. Fortunately we have dynamic programming algorithms, like Needleman-Wunch algorithm, which can calculate $B_n(\delta)$ with complexity $O(n^2)$ (see, e.g., [CZ] page 39).

[1] In this case we are also counting different permutations of indels that correspond to the same alignment.

Lemma 2.5 (see, e.g., [LMT] Claim 2.2)**.** *For every $\delta \in \mathbb{R}$, one-sided derivatives are*

$$B'_n(\delta_-) = \lim_{\Delta\delta \to 0-} \frac{B_n(\delta + \Delta\delta) - B_n(\delta)}{\Delta\delta} = \underline{q}_n(\delta),$$

$$B'_n(\delta_+) = \lim_{\Delta\delta \to 0+} \frac{B_n(\delta + \Delta\delta) - B_n(\delta)}{\Delta\delta} = \overline{q}_n(\delta).$$

Corollary 2.6. *Functions $\overline{q}_n$ and $\underline{q}_n$ satisfy the following properties:*

1. *Both functions are piecewise constant and non-decreasing.*
2. *Function $\overline{q}_n$ is right continuous and $\underline{q}_n$ is left continuous.*
3. *Following equalities hold:*

$$\lim_{\delta \to \infty} \overline{q}_n(\delta) = \lim_{\delta \to \infty} \underline{q}_n(\delta) = 1,$$

$$\lim_{\delta \to -\infty} \overline{q}_n(\delta) = \lim_{\delta \to -\infty} \underline{q}_n(\delta) = 0.$$

Proof. We can conclude from Lemma 2.2 that B_n is not differentiable only in finitely many values, since there is finite amount of possible alignments. Let $\delta_1 < \delta_2 < \ldots < \delta_K$ be such values. Hence, for every

$$(a, b) \in \{(-\infty, \delta_1,), (\delta_2, \delta_3), \ldots, (\delta_{K-1}, \delta_K), (\delta_K, \infty)\},$$

function $B_n\big|_{(a,b)}$ is linear and

$$B_n(\delta) = t(1 - q) + \delta q,$$

for all $\delta \in (a, b)$ and fixed $(t, q) \in \mathcal{O}_n(\delta)$. Therefore $\mathcal{O}_n(\delta)$ is a singleton for $\delta \in (a, b)$ and we can conclude that

$$\overline{q}_n\Big|_{(a,b)} = \underline{q}_n\Big|_{(a,b)}.$$

Since

$$B'_n\Big|_{(a,b)}(\delta) = q,$$

$\overline{q}_n, \underline{q}_n$ are piecewise constant in $\mathbb{R}$. Functions $\overline{q}_n$ and $\underline{q}_n$ are non-decreasing, since increasing gap price, we increase the number of gaps in the optimal alignments and the proportion of gaps cannot decrease.

Next, we will prove the second property. Functions $\overline{q}_n$ and $\underline{q}_n$ are left and right continuous in every interval

$$(a, b) \in \{(-\infty, \delta_1,), (\delta_2, \delta_3), \ldots, (\delta_{K-1}, \delta_K), (\delta_K, \infty)\}.$$

Therefore, we only need to show that

$$\lim_{\delta \to \delta_i +} \overline{q}_n(\delta) = \overline{q}_n(\delta_i),$$

$$\lim_{\delta \to \delta_i -} \underline{q}_n(\delta) = \underline{q}_n(\delta_i),$$

for all $i \in \{1, 2, \ldots, K\}$. Let us fix $i \in \{1, 2, \ldots, K-1\}$, such that

$$\lim_{\delta \to \delta_i +} \overline{q}_n(\delta) = \overline{q}_n(\delta_i + \varepsilon) =: q,$$

for all $\varepsilon \in (0, \delta_{i+1} - \delta_i)$ (for $i = K$ we take $\varepsilon \in (0, \infty)$) and denote $t := t_n(\delta_i + \varepsilon)$. We know that

$$B_n(\delta_i + \varepsilon) = t(1 - q) + q(\delta_i + \varepsilon)$$

and using B_n continuity, we can conclude from

$$\lim_{\varepsilon \to 0+} B_n(\delta_i + \varepsilon) = B_n(\delta_i),$$

that $(t, q) \in \mathcal{O}_n(\delta_i)$. Since $\overline{q}_n$ is non-decreasing, we get $q = \overline{q}_n(\delta_i)$, which means that $\overline{q}_n$ is right continuous.
Proof for left-continuity is analogous.

Lastly, we will prove the third property. If $\delta > F$, the optimal alignment consists only of gaps and

$$\lim_{\delta \to \infty} \overline{q}_n(\delta) = \lim_{\delta \to \infty} \underline{q}_n(\delta) = 1.$$

If $\delta < -nF$, each alignment with at least one gap is negative. Therefore the optimal alignment has no gaps and

$$\lim_{\delta \to -\infty} \overline{q}_n(\delta) = \lim_{\delta \to -\infty} \underline{q}_n(\delta) = 0.$$

■

Lemma 2.7. *There is a constant $\delta_0 \in \mathbb{R}$, such that B_n is strictly increasing in (δ_0, ∞) and*

$$\delta_0 = \min\{\delta \colon \overline{q}_n(\delta) > 0\}.$$

Proof. Let us assume $B_n(\delta_1) < B_n(\delta_2)$ for $\delta_1 < \delta_2$ and B_n is not strictly increasing in (δ_2, ∞). In other words, there exists $\delta_3 > \delta_2$, such that $B_n(\delta_2) = B_n(\delta_3)$. Let us fix

$$\lambda = \frac{\delta_3 - \delta_2}{\delta_3 - \delta_1} \in (0, 1),$$

then $\delta_2 = \lambda\delta_1 + (1 - \lambda)\delta_3$ and

$$\begin{aligned} B_n(\delta_2) &= B_n(\lambda\delta_1 + (1 - \lambda)\delta_3) \\ &> B_n(\delta_1) \\ &= \lambda B_n(\delta_1) + (1 - \lambda)B_n(\delta_1) \\ &> \lambda B_n(\delta_1) + (1 - \lambda)B_n(\delta_3), \end{aligned}$$

which is a contradiction with B_n being convex. Therefore there exists a constant δ_0, such that B_n is strictly increasing in (δ_0, ∞).

For all $\delta_1', \delta_2' \in (\min\{\delta \colon \overline{q}_n(\delta) > 0\}, \infty)$, such that $\delta_1' < \delta_2'$

$$\begin{aligned} B_n(\delta_1') &= \max_{(t,q)\in\overline{\mathcal{O}}_n'} (t(1-q) + \delta_1' q) \\ &= t^*(1-q^*) + \delta_1' q^* \\ &< t^*(1-q^*) + \delta_2' q^* \qquad (2.1) \\ &\leq \max_{(t,q)\in\overline{\mathcal{O}}_n'} (t(1-q) + \delta_2' q) \\ &= B_n(\delta_2'), \end{aligned}$$

where $(t^*, q^*) \in \overline{\mathcal{O}}_n'$ is an optimal pair and (2.1) follows from $q^* \neq 0$, which we can always find, since $\overline{q}_n(\delta_1') > 0$ and $\overline{q}_n(\delta_2') > 0$. It should be noted that the existence of

$$\min\{\delta \colon \overline{q}_n(\delta) > 0\}$$

comes from $\overline{q}_n$ right-continuouty. ∎

Example 2.8. Let us fix the alphabet as $\mathbb{A} = \{a, b, c\}$ and sequences

$$\begin{aligned} x &= (a, c, b, b, c, b, a, a, a, c, b, c, b, b, a, a, c, b, a, c), \\ y &= (c, b, b, b, a, c, a, b, a, c, a, b, c, b, b, c, a, a, c, b). \end{aligned}$$

The score function is appointed as

$$\begin{aligned} &S(a,b) = 0, && S(a,c) = 0, && S(b,c) = 0, \\ &S(a,a) = 1, && S(b,b) = 2, && S(c,c) = 3. \end{aligned}$$

Next, we will sketch the graph of B_{20}.

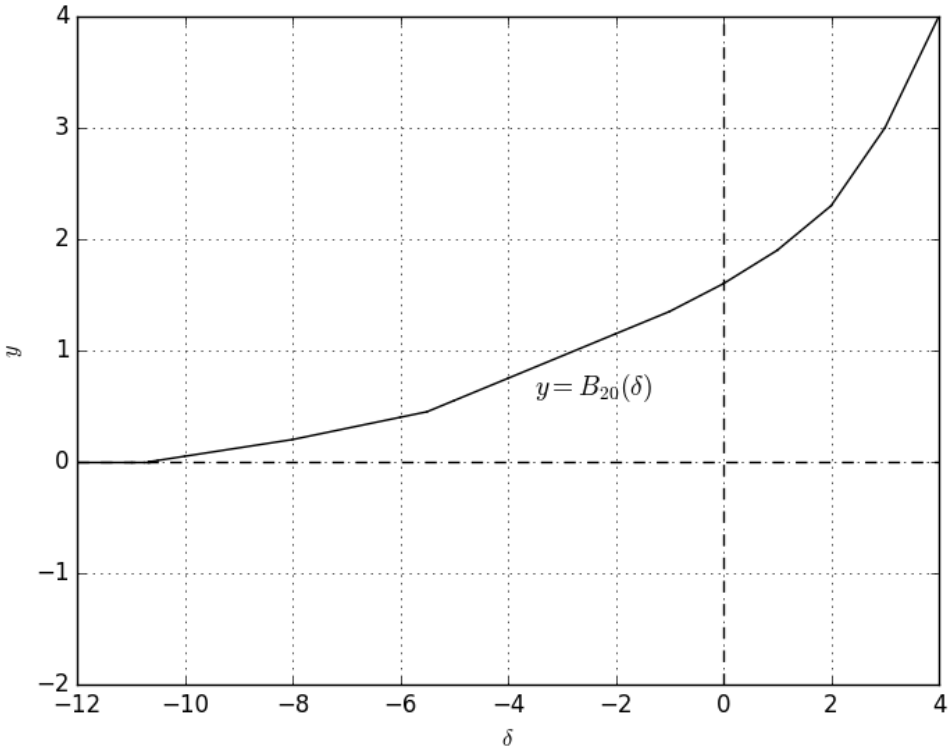

Figure 2.9: Graph of the function B_{20}

It is easier to notice each value, in which optimal alignments change, if we sketch the graph of q_{20}. In each constant piece of q_{20}, optimal alignment stays the same and the change in the value of B_{20} comes only from the gap price δ.

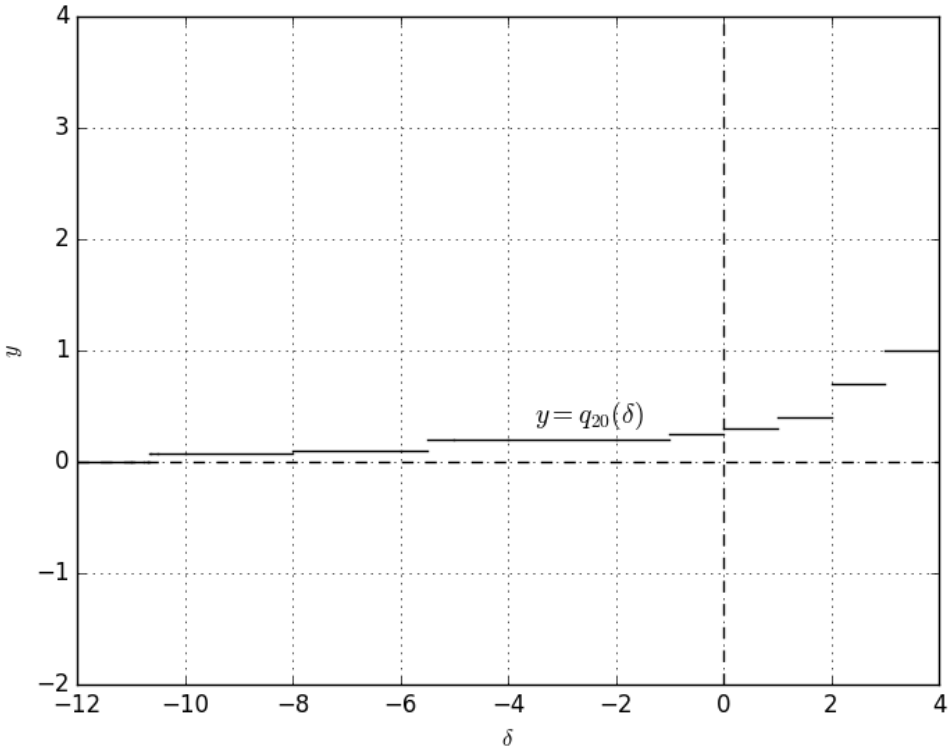

Figure 2.10: Graph of the function q_{20}

We can see from figure 2.10, that $\delta_0 \approx -10.5$, which is defined in Lemma 2.7.

Chapter 3

Proportion of Mismatches

This chapter is devoted to introducing the theory that covers the effect of varying score of mismatch has on the optimal alignment score. In addition, we also define an optimality region and illustrate the theory with numerous examples. We show that many analogous results hold, to varying the gap price, that we described in the previous chapter.

In this chapter, we fix the alphabet as $\mathbb{A} = \{a, b\}$ and the score function $S\colon \mathbb{A}\times\mathbb{A} \to \mathbb{R}$ as

$$S(a,a) = S(b,b) = 1, \qquad S(a,b) = S(b,a) = \zeta.$$

We have decided to restrict our theory by fixing the score of matched letters and letting the score of mismatched letters vary. This can be thought of as a first step of generalization, extension to only varying the gap price. Theory can be developed further by considering non-binary alphabets and letting all parameters of the score function S vary, in addition to the gap price.

Definition 3.1. Let k_{match} denote the number of matches. The number

$$p(\pi,\mu) := \frac{k - k_{\text{match}}}{n},$$

where k is the number of aligned letters, is called the **proportion of mismatches**.

Proportion of mismatches depends on the alignment (π, μ). Generally, we abandon the alignment in order to simplify notation. Score of the alignment can written as

$$\begin{aligned} U^{\delta,\zeta}_{(\pi,\mu)}(x,y) &= n((1-p-q) + \zeta p + \delta q) \\ &= n(1 + p(\zeta - 1) + q(\delta - 1)). \end{aligned}$$

We define the two-variable function $B_n\colon \mathbb{R}^2 \to \mathbb{R}$ as

$$B_n(\zeta,\delta) = \max_{(p,q)\in\overline{\mathcal{O}}'_n} (1 + p(\zeta - 1) + q(\delta - 1)),$$

The next example displays the behaviour of the two-variable function B_n and proportions of gaps and mismatches. It turns out that the behaviour is similar to the one-variable case, that we already covered in Example 1.8 and Example 2.8.

Example 3.2. Let us fix sequences

$$x = (a, b, a, a, b, a, b, b, b, a, b, b),$$
$$y = (b, a, b, b, b, a, a, b, a, b, b, a).$$

Next, we will fix $\zeta^* = 0.5$ and sketch graphs of the one-variable functions $\delta \mapsto B_{12}(\zeta^*, \delta)$ and $\delta \mapsto q_{12}(\zeta^*, \delta)$.

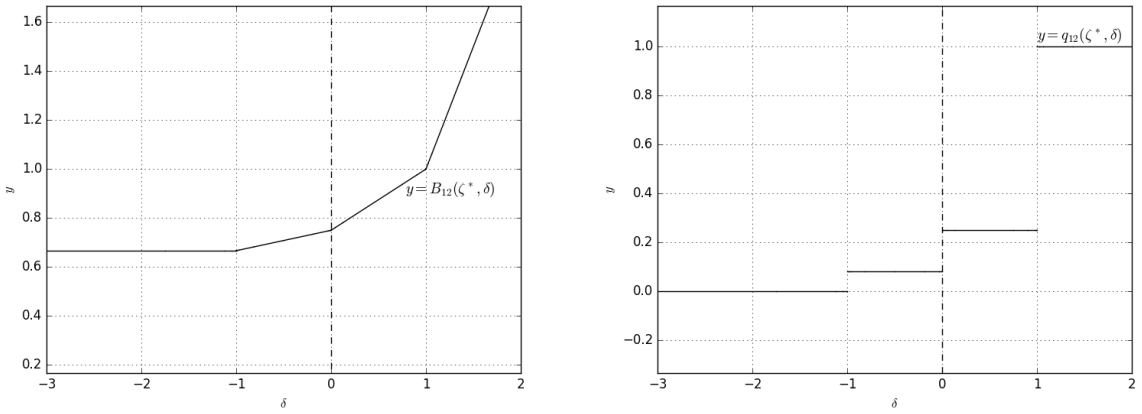

Figure 3.3: Graphs of $\delta \mapsto B_{12}(\zeta^*, \delta)$ and $\delta \mapsto q_{12}(\zeta^*, \delta)$

Let us fix $\delta^* = 0.25$. We can sketch graphs of the functions $\zeta \mapsto B_{12}(\zeta, \delta^*)$ and $\zeta \mapsto p_{12}(\zeta, \delta^*)$.

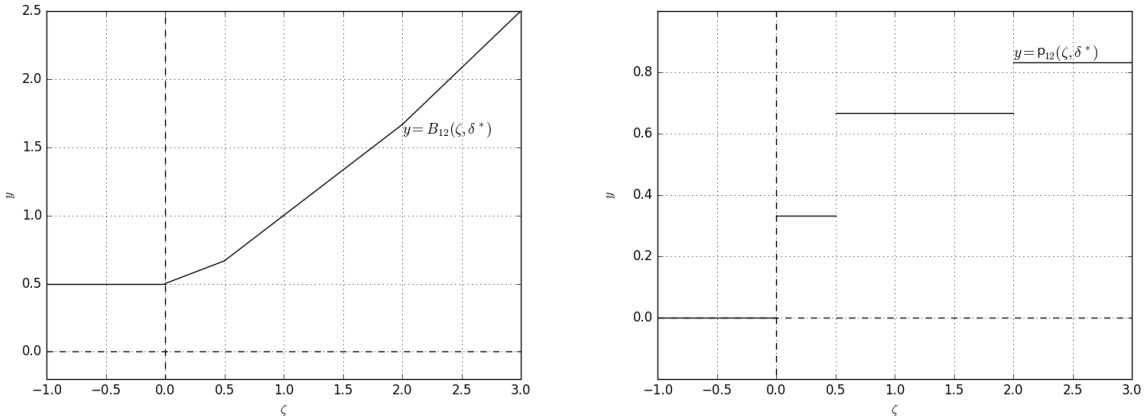

Figure 3.4: Graphs of $\zeta \mapsto B_{12}(\zeta, \delta^*)$ and $\zeta \mapsto p_{12}(\zeta, \delta^*)$

It turns out that considering B_n, considered as a two-variable function, preserves some fundamental properties from the one-variable case.

Lemma 3.5. *Two-variable function $B_n\colon \mathbb{R}^2 \to \mathbb{R}$ is continuous.*

Proof. Let us fix sequences $\zeta_i \to \zeta$ and $\delta_i \to \delta$. Then for every $(\zeta, \delta) \in \mathbb{R}^2$

$$\begin{aligned}\lim_{i\to\infty} B_n(\zeta_i, \delta_i) &= \lim_{i\to\infty} \max_{(p,q)\in\overline{\mathcal{O}}'_n} (1 + p(\zeta_i - 1) + q(\delta_i - 1)) \\ &= \max_{(p,q)\in\overline{\mathcal{O}}'_n} \lim_{i\to\infty} (1 + p(\zeta_i - 1) + q(\delta_i - 1)) \qquad (3.1) \\ &= \max_{(p,q)\in\overline{\mathcal{O}}'_n} (1 + p(\zeta - 1) + q(\delta - 1)) \\ &= B_n(\zeta, \delta).\end{aligned}$$

Since the maximum is taken over finite number of possible alignments, function B_n is defined as a maximum of finite number of affine functions. That justifies the equality (3.1). ∎

Definition 3.6. Set of 2-tuples $\mathcal{O}_n(\zeta, \delta)$ is called the **set of optimal 2-tuples**, corresponding to sequences x, y and parameters n, ζ, δ, if

$$(p, q) \in \mathcal{O}_n(\zeta, \delta) \Leftrightarrow B_n(\zeta, \delta) = 1 + p(\zeta - 1) + q(\delta - 1),$$

where p, q corresponds to an alignment of x and y.

Let us denote

$$\overline{p}_n(\zeta, \delta) := \max\{p\colon (p, q) \in \mathcal{O}_n(\zeta, \delta)\},$$
$$\underline{p}_n(\zeta, \delta) := \min\{p\colon (p, q) \in \mathcal{O}_n(\zeta, \delta)\}.$$

For convenience, we write

$$p_n(\zeta, \delta) := \overline{p}_n(\zeta, \delta) = \underline{p}_n(\zeta, \delta).$$

3.1 Optimality Regions

It turns out that decomposition of $\mathcal{O}_n(\zeta, \delta)$ is surprisingly structured. Our main area of interest is in values $(\zeta, \delta) \in \mathbb{R}^2$, where $\mathcal{O}_n(\zeta, \delta)$ is not a singleton, hence the optimal alignment is not uniquely determined at (ζ, δ). We refer to [GBN] page 314 for the structure of sets, where $\mathcal{O}_n(\zeta, \delta)$ is a singleton. They are semi-infinite cones.

Definition 3.7. Each convex polygon $\{(\zeta, \delta) \in \mathbb{R}^2\colon \mathcal{O}_n(\zeta, \delta) \text{ is a singleton}\}$ is called a **optimality region**.

Lemma 3.8 (see, e.g., [GBN], page 319)**.** *The number of optimality regions is bounded by $O(n^{2/3})$.*

Lemma 3.9 (see, e.g., [BSS], Lemma 3 and note that $\beta = -\frac{1}{2}\delta$ and $\alpha = -\zeta$)**.** *All optimality regions in*

$$\{(\zeta, \delta) \in \mathbb{R}^2\colon \zeta < 0,\ \delta < 0\}$$

are semi-infinite cones and are delimited by the coordinates axes or by lines of the form $-\frac{1}{2}\delta = c - \left(c + \frac{1}{2}\right)\zeta$ for some constant $c \in \mathbb{R}$.

3.2 Properties of the Two-variable Function B_n

The next lemma is generalization of Claim 2.1 from [BSS] for the two-variable case.

Lemma 3.10. *The function B_n is piecewise affine, convex and non-decreasing in relation to each of its variables.*

Proof. Using Lemma 3.8, we know that the optimal alignment changes only in finitely many optimality regions. Let us denote them as $V_1, V_2, \ldots, V_j$, $j \in \mathbb{N}$. Then for each $i \in \{1, 2, \ldots, j\}$

$$B_n\big|_{V_i}(\zeta, \delta) = 1 + p(\zeta - 1) + q(\delta - 1),$$

for (p, q) corresponding to the optimality region V_i. If $\lambda \in \mathbb{R}$ and

$$(\zeta_1, \delta_1), (\zeta_1, \delta_2), (\zeta_1 + \lambda\zeta_2, \delta_1 + \lambda\delta_2) \in V_i,$$

then B_n is affine, since

$$\begin{aligned} B_n(\zeta_1 + \lambda\zeta_2, \delta_1 + \lambda\delta_2) + \lambda(1 + p + q) &= 1 + p(\zeta_1 + \lambda\zeta_2 - 1) + q(\delta_1 + \lambda\delta_2 - 1) \\ &\quad + \lambda(1 + p + q) \\ &= 1 + p(\zeta_1 - 1) + q(\delta_1 - 1) \\ &\quad + \lambda(1 + p(\zeta_2 - 1) + q(\delta_2 - 1)) \\ &= B_n(\zeta_1, \delta_1) + \lambda B_n(\zeta_2, \delta_2). \end{aligned}$$

Since all optimality regions are convex, B_n is piecewise affine in each of them.

For the convexity, let us fix $\lambda \in (0, 1)$, $(\zeta_1, \delta_1), (\zeta_2, \delta_2) \in \mathbb{R}^2$ and denote

$$(\zeta^*, \delta^*) := \lambda(\zeta_1, \delta_1) + (1 - \lambda)(\zeta_2, \delta_2).$$

For a fixed $(p^*, q^*) \in \mathcal{O}_n(\zeta^*, \delta^*)$, we have

$$\begin{aligned} B_n(\zeta^*, \delta^*) &= B_n(\lambda(\zeta_1, \delta_1) + (1 - \lambda)(\zeta_2, \delta_2)) \\ &= 1 + p^*(\lambda\zeta_1 + (1 - \lambda)\zeta_2 - 1) + q^*(\lambda\delta_1 + (1 - \lambda)\delta_2 - 1). \end{aligned} \tag{3.2}$$

Since

$$B_n(\zeta_1, \delta_1) = \max_{(p,q) \in \overline{\mathcal{O}}'_n} (1 + p(\zeta_1 - 1) + q(\delta_1 - 1)) \geq 1 + p^*(\zeta_1 - 1) + q^*(\delta_1 - 1),$$

$$B_n(\zeta_2, \delta_2) = \max_{(p,q) \in \overline{\mathcal{O}}'_n} (1 + p(\zeta_2 - 1) + q(\delta_2 - 1)) \geq 1 + p^*(\zeta_2 - 1) + q^*(\delta_2 - 1).$$

we get from (3.2)

$$B_n(\zeta^*, \delta^*) \leq \lambda B_n(\zeta_2, \delta_1) + (1 - \lambda) B_n(\zeta_2, \delta_2).$$

Therefore B_n is convex.

For the non-decreasing property, we fix $\zeta^* \in \mathbb{R}$. For all $\delta_0 < \delta_1$

$$\begin{aligned} B_n(\zeta^*, \delta_0) &= \max_{(p,q)\in\overline{\mathcal{O}}_n'} (1 + p(\zeta^* - 1) + q(\delta_0 - 1)) \\ &= 1 + p^*(\zeta^* - 1) + q^*(\delta_0 - 1) \\ &\leq 1 + p^*(\zeta^* - 1) + q^*(\delta_1 - 1) \\ &\leq \max_{(p,q)\in\overline{\mathcal{O}}_n'} (1 + p(\zeta^* - 1) + q(\delta_1 - 1)) \\ &= B_n(\zeta^*, \delta_1), \end{aligned}$$

where $(p^*, q^*) \in \mathcal{O}_n(\zeta^*, \delta_0)$. Similarly, when fixing $\delta^* \in \mathbb{R}$, for all $\zeta_0 < \zeta_1$

$$\begin{aligned} B_n(\zeta_0, \delta^*) &= \max_{(p,q)\in\overline{\mathcal{O}}_n'} (1 + p(\zeta_1 - 1) + q(\delta^* - 1)) \\ &= 1 + p^{**}(\zeta_0 - 1) + q^{**}(\delta^* - 1) \\ &\leq 1 + p^{**}(\zeta_1 - 1) + q^{**}(\delta^* - 1) \\ &\leq \max_{(p,q)\in\overline{\mathcal{O}}_n'} (1 + p(\zeta_1 - 1) + q(\delta^* - 1)) \\ &= B_n(\zeta_1, \delta^*), \end{aligned}$$

where $(p^{**}, q^{**}) \in \mathcal{O}_n(\zeta_0, \delta^*)$. ∎

Example 3.11. Let us fix sequences

$$\begin{aligned} x &= (a, a, b, a, b, b, a, b, a, a), \\ y &= (b, a, b, a, a, a, b, a, a, b). \end{aligned}$$

Let us fix $\delta^* = -20$ and sketch the graph of $\zeta \mapsto B_{10}(\zeta, \delta^*)$.

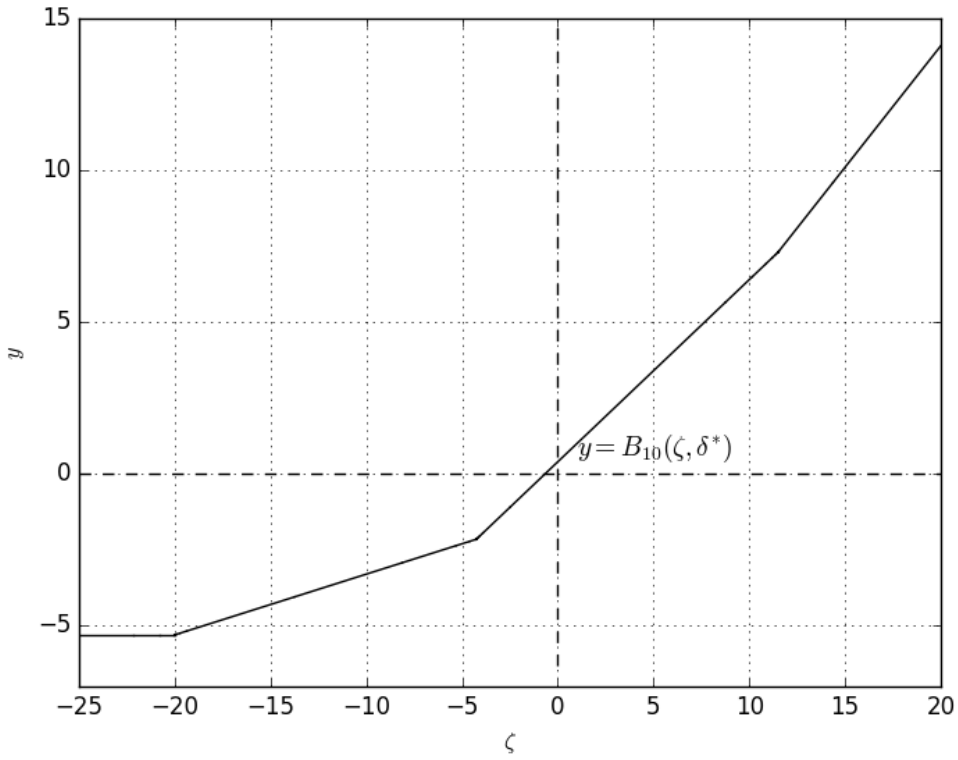

Figure 3.12: Graph of $\zeta \mapsto B_{10}(\zeta, \delta^*)$

We can approximate the values $\zeta_1, \zeta_2, \zeta_3 \in \mathbb{R}$, at which the optimal alignment changes and B_n is not differentiable:

$$(\zeta_1, \delta^*) \approx (-20, -20), \quad (\zeta_2, \delta^*) \approx (-4, -20), \quad (\zeta_3, \delta^*) \approx (12, -20).$$

Using these points on the plane and since Lemma 3.9 gives us the general form of all lines, we can sketch optimality regions. Since each line goes through $(1, 1)$ (see, e.g., [GBN], Theorem 4.1.), we can use approximated (ζ_1, δ^*) and (ζ_2, δ^*) and draw lines through those two points, to sketch optimality regions in the lower left quarter of the plane.

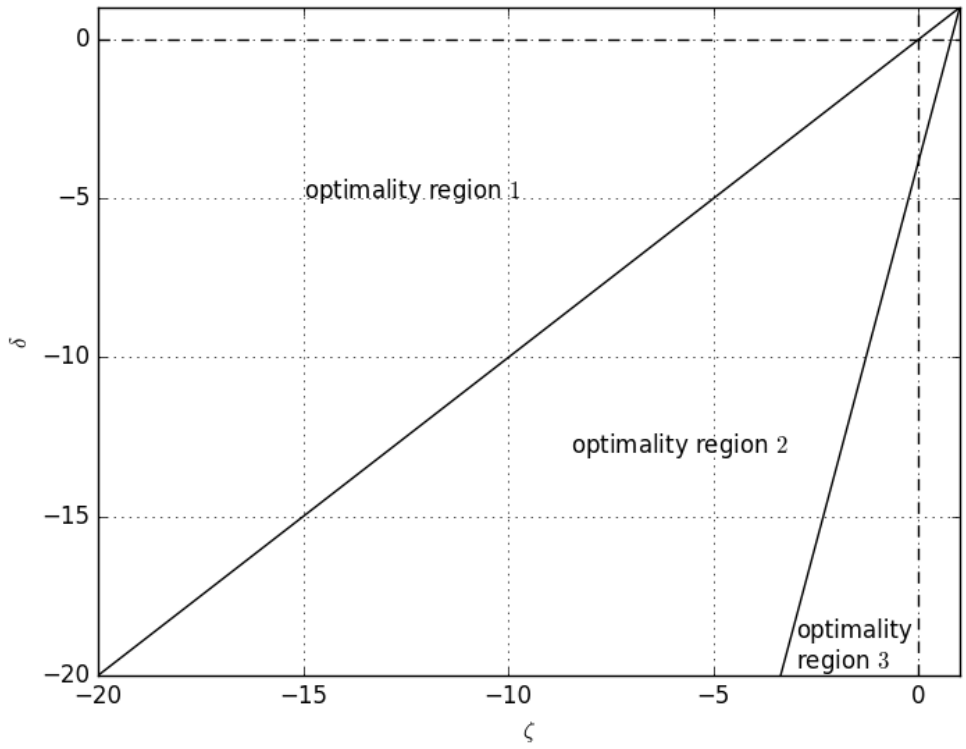

Figure 3.13: Optimality regions

The next lemma is a generalization of Claim 2.2 from [LMT], for the two-variable case. We denote one-sided partial derivatives as

$$\begin{aligned}
\frac{\partial B_n(\zeta_-,\delta)}{\partial\zeta} &:= \lim_{h\to 0-} \frac{B_n(\zeta+h,\delta)-B_n(\zeta,\delta)}{h},\\
\frac{\partial B_n(\zeta_+,\delta)}{\partial\zeta} &:= \lim_{h\to 0+} \frac{B_n(\zeta+h,\delta)-B_n(\zeta,\delta)}{h},\\
\frac{\partial B_n(\zeta,\delta_-)}{\partial\delta} &:= \lim_{h\to 0-} \frac{B_n(\zeta,\delta+h)-B_n(\zeta,\delta)}{h},\\
\frac{\partial B_n(\zeta,\delta_+)}{\partial\delta} &:= \lim_{h\to 0+} \frac{B_n(\zeta,\delta+h)-B_n(\zeta,\delta)}{h}.
\end{aligned}$$

Lemma 3.14. *For all $(\zeta,\delta)\in\mathbb{R}^2$, one-sided derivatives of B_n are*

a) $\dfrac{\partial B_n(\zeta_-,\delta)}{\partial\zeta} = \underline{p}_n(\zeta,\delta),$ *b)* $\dfrac{\partial B_n(\zeta_+,\delta)}{\partial\zeta} = \overline{p}_n(\zeta,\delta),$

c) $\dfrac{\partial B_n(\zeta,\delta_-)}{\partial\delta} = \underline{q}_n(\zeta,\delta),$ *d)* $\dfrac{\partial B_n(\zeta,\delta_+)}{\partial\delta} = \overline{q}_n(\zeta,\delta).$

Therefore $\mathcal{O}_n(\zeta,\delta)$ is a singleton, if function B_n has all its partial derivatives at (ζ,δ).

Proof. We start by proving equation *a)*. Let us fix $(\zeta^*,\delta^*)\in\mathbb{R}^2$, $s_1>0$ and $(p^*,q^*)\in\mathcal{O}_n(\zeta^*,\delta^*)$, thus

$$\begin{aligned} B_n(\zeta^*,\delta^*+s_1) &= \max_{(p,q)\in\overline{\mathcal{O}}'_n}(1+p(\zeta^*-1)+q(\delta^*+s_1-1)) \\ &\geq 1+p^*(\zeta^*-1)+q^*(\delta^*+s_1-1) \\ &= B_n(\zeta^*,\delta^*)+s_1q^*, \end{aligned} \tag{3.3}$$

similarly we get

$$B_n(\zeta^*,\delta^*-s_1)\geq 1+p^*(\zeta^*-1)+q^*(\delta^*-s_1-1)=B_n(\zeta^*,\delta^*)-s_1q^*.$$

Hence

$$\frac{B_n(\zeta^*,\delta^*)-B_n(\zeta^*,\delta^*-s_1)}{s_1}\leq q^*\leq\frac{B_n(\zeta^*,\delta^*+s_1)-B_n(\zeta^*,\delta^*)}{s_1}.$$

Letting $s_1\to 0+$ and using the Sandwich Theorem, we get

$$\frac{\partial B_n(\zeta^*,\delta^*_-)}{\partial\delta}\leq q_*\leq\frac{\partial B_n(\zeta^*,\delta^*_+)}{\partial\delta}. \tag{3.4}$$

Existence of one-sided partial derivatives comes from B_n being convex (see, e.g., [Roc] Theorem 23.1). Since (3.4) holds for any optimal pair $(p,q)\in\mathcal{O}_n(\zeta^*,\delta^*)$, we get

$$\frac{\partial B_n(\zeta^*,\delta^*_-)}{\partial\delta}\leq\underline{q}_n(\zeta^*,\delta^*)\leq\overline{q}_n(\zeta^*,\delta^*)\leq\frac{\partial B_n(\zeta^*,\delta^*_+)}{\partial\delta}.$$

Therefore the partial derivative exists at (ζ^*,δ^*), if $\mathcal{O}_n(\zeta^*,\delta^*)$ is a singleton. It is enough to show that there exists $(p,q)\in\mathcal{O}_n(\zeta^*,\delta^*)$, such that

$$\frac{\partial B_n(\zeta^*,\delta^*_-)}{\partial\delta}=\underline{q}_n(\zeta^*,\delta^*). \tag{3.5}$$

Since B_n is piecewise affine, for every small $\varepsilon_1>0$, the optimal alignment does not change at $(\zeta^*,\delta^*-\varepsilon_1)$ and B_n is differentiable. Let us denote

$$q_1:=q(\zeta^*,\delta^*-\varepsilon_2)=\frac{\partial B_n(\zeta^*,\delta^*_-)}{\partial\delta},$$

$$p_1:=p(\zeta^*,\delta^*-\varepsilon_1),$$

Thus, for every $\varepsilon_1>0$ small enough there exists $(p_1,q_1)\in\mathcal{O}_n(\zeta^*,\delta^*-\varepsilon_1)$, such that

$$B_n(\zeta^*,\delta^*-\varepsilon_1)=1+p_1(\zeta^*-1)+q_1(\delta^*-\varepsilon_1-1).$$

From Lemma 3.5, we know that B_n is a continuous function, hence

$$\lim_{\varepsilon_1\to 0+}B_n(\zeta^*,\delta^*-\varepsilon_1)=B_n(\zeta^*,\delta^*)=1+p_1(\zeta^*-1)+q_1(\delta^*-1).$$

Therefore $(p_1, q_1) \in \mathcal{O}_n(\zeta^*, \delta^*)$, hence (3.5) holds. With similar arguments one can show *b)*. Next, we will prove *c)*. Let us fix $s_2 > 0$. With similar arguments as in (3.3), we get

$$B_n(\zeta^* + s_2, \delta^*) \geq 1 + p^*(\zeta^* + s_2 - 1) + q^*(\delta^* - 1) = B_n(\zeta^*, \delta^*) + s_2 p^*,$$

$$B_n(\zeta^* - s_2, \delta^*) \geq 1 + p^*(\zeta^* - s_2 - 1) + q^*(\delta^* - 1) = B_n(\zeta^*, \delta^*) - s_2 p^*.$$

Hence

$$\frac{B_n(\zeta^*, \delta^*) - B_n(\zeta^* - s_2, \delta^*)}{s_2} \leq p^* \leq \frac{B_n(\zeta^* + s_2, \delta^*) - B_n(\zeta^*, \delta^*)}{s_2}.$$

Letting $s_2 \to 0+$ and using the Sandwich Theorem, we get

$$\frac{\partial B_n(\zeta^*_-, \delta^*)}{\partial \zeta} \leq p_* \leq \frac{\partial B_n(\zeta^*_+, \delta^*)}{\partial \zeta}. \tag{3.6}$$

Existence of one-sided partial derivatives comes from B_n being convex (see, e.g., [Roc] Theorem 23.1). Since (3.6) holds for any optimal pair $(p, q) \in \mathcal{O}_n(\zeta^*, \delta^*)$, we get

$$\frac{\partial B_n(\zeta^*_-, \delta^*)}{\partial \zeta} \leq \underline{p}_n(\zeta^*, \delta^*) \leq \overline{p}_n(\zeta^*, \delta^*) \leq \frac{\partial B_n(\zeta^*_+, \delta^*)}{\partial \zeta}.$$

Therefore the partial derivative exists at (ζ^*, δ^*), if $\mathcal{O}_n(\zeta^*, \delta^*)$ is a singleton. It is enough to show that there exists $(p, q) \in \mathcal{O}_n(\zeta^*, \delta^*)$, such that

$$\frac{\partial B_n(\zeta^*_-, \delta^*)}{\partial \zeta} = \underline{p}_n(\zeta^*, \delta^*). \tag{3.7}$$

Since B_n is piecewise affine, for every small $\varepsilon_2 > 0$, the optimal alignment does not change at $(\zeta^* - \varepsilon_2, \delta^*)$ and B_n is differentiable. Let us denote

$$q_2 := q(\zeta^* - \varepsilon_2, \delta^*),$$

$$p_2 := p(\zeta^* - \varepsilon_2, \delta^*) = \frac{\partial B_n(\zeta^*_-, \delta^*)}{\partial \zeta},$$

Thus, for every $\varepsilon_2 > 0$ small enough there exists $(p_2, q_2) \in \mathcal{O}_n(\zeta^* - \varepsilon_2, \delta^*)$, such that

$$B_n(\zeta^* - \varepsilon_2, \delta^*) = 1 + p_2(\zeta^* - \varepsilon_2 - 1) + q_2(\delta^* - 1).$$

From Lemma 3.5, we know that B_n is a continuous function, hence

$$\lim_{\varepsilon_2 \to 0+} B_n(\zeta^* - \varepsilon_2, \delta^*) = B_n(\zeta^*, \delta^*) = 1 + p_2(\zeta^* - 1) + q_2(\delta^* - 1).$$

Therefore $(p_2, q_2) \in \mathcal{O}_n(\zeta^*, \delta^*)$, hence (3.7) holds. With similar arguments one can show *d)*. ∎

Example 3.15. In this example we demonstrate the geometric interpretation of $\underline{p}_n$ and $\overline{p}_n$. Let us fix sequences

$$x = (a, b, a, a, b, a, b, b, b, a, a, b, a, b, a, a, b, a, a, b),$$
$$y = (b, a, b, b, a, b, a, a, b, a, b, b, b, a, b, a, a, b, a, b)$$

and the gap price as $\delta^* = 0.5$. Next, we will sketch graphs of the one-variable functions $\zeta \mapsto p_{20}(\zeta, \delta^*)$ and $\zeta \mapsto B_{20}(\zeta, \delta^*)$.

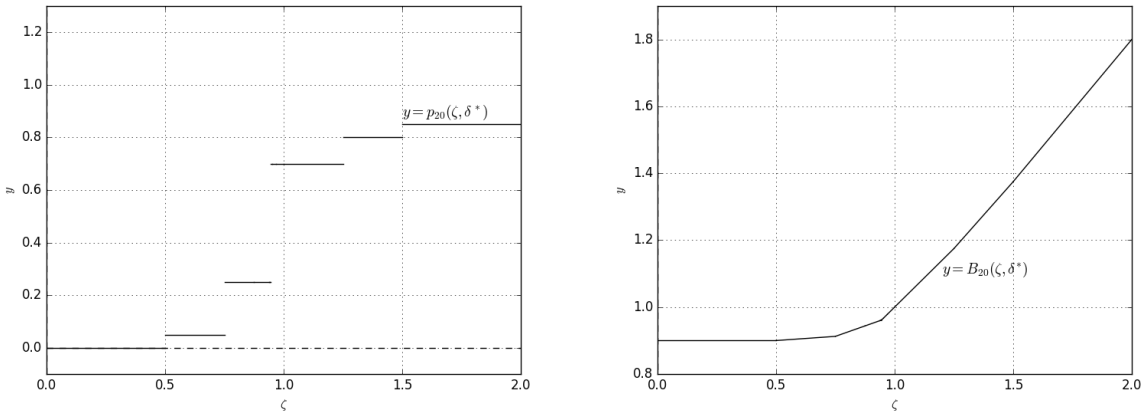

Figure 3.16: Graphs of the functions $\zeta \mapsto p_{20}(\zeta, \delta^*)$ and $\zeta \mapsto B_{20}(\zeta, \delta^*)$

We can approximate the values $\zeta \in \mathbb{R}$, at which $\underline{p}_{20}(\zeta, \delta^*) \neq \overline{p}_{20}(\zeta, \delta^*)$, from the graph in Figure 3.16:

$$\begin{aligned}
\overline{p}_{20}(0.5, \delta^*) &\approx 0.05, & \underline{p}_{20}(0.5, \delta^*) &\approx 0,\\
\overline{p}_{20}(0.75, \delta^*) &\approx 0.25, & \underline{p}_{20}(0.75, \delta^*) &\approx 0.05,\\
\overline{p}_{20}(0.9, \delta^*) &\approx 0.7, & \underline{p}_{20}(0.9, \delta^*) &\approx 0.25,\\
\overline{p}_{20}(1.25, \delta^*) &\approx 0.8, & \underline{p}_{20}(1.25, \delta^*) &\approx 0.7,\\
\overline{p}_{20}(1.5, \delta^*) &\approx 0.85, & \underline{p}_{20}(1.5, \delta^*) &\approx 0.8.
\end{aligned}$$

Fixing $\delta^{**} = -2$, gives us the following graphs:

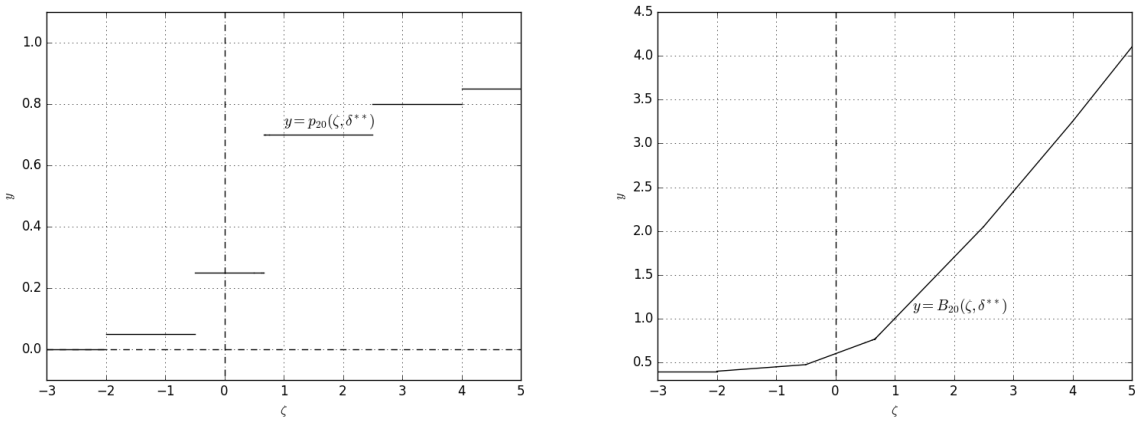

Figure 3.17: Graphs of the functions $\zeta \mapsto p_{20}(\zeta, \delta^{**})$ and $\zeta \mapsto B_{20}(\zeta, \delta^{**})$

We can approximate the values $\zeta \in \mathbb{R}$, at which $\underline{p}_{20}(\zeta, \delta^{**}) \neq \overline{p}_{20}(\zeta, \delta^{**})$ from the graph in Figure 3.17:

$$\begin{aligned} \overline{p}_{20}(-2, \delta^{**}) &\approx 0.05, & \underline{p}_{20}(-2, \delta^{**}) &\approx 0, \\ \overline{p}_{20}(-0.5, \delta^{**}) &\approx 0.25, & \underline{p}_{20}(-0.5, \delta^{**}) &\approx 0.05, \\ \overline{p}_{20}(0.75, \delta^{**}) &\approx 0.7, & \underline{p}_{20}(0.75, \delta^{**}) &\approx 0.25, \\ \overline{p}_{20}(2.5, \delta^{**}) &\approx 0.8, & \underline{p}_{20}(2.5, \delta^{**}) &\approx 0.7, \\ \overline{p}_{20}(4, \delta^{**}) &\approx 0.85, & \underline{p}_{20}(4, \delta^{**}) &\approx 0.8. \end{aligned}$$

Chapter 4

The Asymptotic Proportion of Mismatches and Gaps

Following [LMT], we prove analogous results for asymptotic proportion of mismatches. We prove that B_n converges almost surely in $\mathbb{R}^2$ and that the limit function b is differentiable almost everywhere, similarly to the one-variable case. The second part of the chapter is devoted to simulations, that illustrate the geometric meaning of previous results in this chapter.

4.1 Properties of the Two-variable Limit Function b

Following theorem and lemma are subsidiary results, which we use to prove almost surely convegence.

Theorem 4.1 (Kingman's Subadditive Ergodic Theorem)**.** (see, e.g., [Lem] Teoreem 6.10)**.** *Let $X_1, X_2, \ldots$ be iid random variables. Let g_n, $n \in \mathbb{N}$ be functions, such that*

$$\mathbf{E}|g_n(X_1, X_2, \ldots)| < \infty, \forall n \in \mathbb{N}$$

and

$$g_{m+v}(X_1, X_2, \ldots) \leq g_m(X_1, X_2, \ldots) + g_v(X_{m+1}, X_{m+2}, \ldots), \forall m, v \in \mathbb{N}.$$

Then there exists a constant $c \in \mathbb{R}$, such that

$$\lim_{n\to\infty} \frac{1}{n} g(X_1, X_2, \ldots) = c, \ \textit{a.s. and in } L_1.$$

Kingman's subadditive ergodic theorem holds under more general conditions, but for this thesis, we have opted for less generalized version.

Lemma 4.2. *For all parameters $(\zeta, \delta) \in \mathbb{R}^2$ and sequences of length $m, v \in \mathbb{N}$, where B_{m+v} is calculated from consequently attaching first m length sequence to the second v length sequence. Following inequality holds*

$$(m+v)B_{m+v}(\zeta,\delta) \geq mB_m(\zeta,\delta) + vB_v(\zeta,\delta).$$

Proof. Let us fix parameters $(\zeta, \delta) \in \mathbb{R}^2$, sequences

$$\begin{aligned} x^m &= (x_1, x_2, \ldots, x_m), \\ y^m &= (y_2, y_2, \ldots, y_m), \\ x^v &= (x_{m+1}, x_{m+2}, \ldots, x_{m+v}), \\ y^v &= (y_{m+1}, y_{m+2}, \ldots, y_{m+v}), \end{aligned}$$

their optimal pairs (p_m, q_m), (p_v, q_v) and number of aligned letters k_m, k_v. Considering sequences

$$\begin{aligned} x &= (x_1, x_2, \ldots, x_{m-1}, x_m, x_{m+1}, \ldots, x_{m+v}), \\ y &= (y_1, y_2, \ldots, y_{m-1}, y_m, y_{m+1}, \ldots, y_{m+v}), \end{aligned}$$

obviously $(p_m + p_v, q_m + q_v) \in \overline{\mathcal{O}}'_{m+v}$ and

$$\begin{aligned} (m+v)B_{m+v}(\zeta,\delta) &= \max_{(p,q)\in\overline{\mathcal{O}}'_n} (k + p\zeta + q\delta) \\ &\geq (k_m + k_v + (p_m + p_v)\zeta + (q_m + q_v)\delta) \\ &= k_m + p_m\zeta + q_m\delta + k_v + p_v\zeta + q_v\delta \\ &= mB_m(\zeta,\delta) + vB_v(\zeta,\delta). \end{aligned}$$

■

Lemma 4.3. *There exists a limit function $b\colon \mathbb{R}^2 \to \mathbb{R}$, such that for all $(\zeta, \delta) \in \mathbb{R}^2$*

$$\lim_{n\to\infty} B_n(\zeta,\delta) = b(\zeta,\delta), \ a.s..$$

Proof. Let us fix parameters $(\zeta, \delta) \in \mathbb{R}^2$ and define functions

$$g_n(x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_n) := -nB_n(\zeta,\delta)^{x_1,x_2,\ldots,x_n;y_1,y_2,\ldots,y_n},$$

for $n \in \mathbb{N}$, where $B_n(\zeta,\delta)^{x_1,x_2,\ldots,x_n;y_1,y_2,\ldots,y_n}$ denotes $B_n(\zeta,\delta)$ for sequences $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$. Lemma 4.2 gives us

$$\begin{aligned} g_{m+v}(x_1, x_2, \ldots x_{m+v}; y_1, y_2, \ldots, y_{m+v}) \leq\ & g_m(x_1, x_2, \ldots, x_m; y_1, y_2, \ldots, y_m) \\ & + g_v(x_{m+1}, x_{m+2}, \ldots, x_{m+v}; \\ & y_{m+1}, y_{m+2}, \ldots, y_{m+v}). \end{aligned}$$

For each $m \in \mathbb{N}$ there are finite amount of sequences $x = (x_1, x_2, \ldots, x_m)$, $y = (y_1, y_2, \ldots, y_m)$ and for all sequences of length m

$$|g_m(x_1, x_2, \ldots, x_m; y_1, y_2, \ldots, y_m)| \leq m(|\zeta| + |\delta| + 1). \tag{4.1}$$

Hence

$$\mathbf{E}|g_m(X_1, X_2, \ldots, X_m; Y_1, Y_2, \ldots, Y_m)| < \infty.$$

Since (X_i, Y_i), $i \in \mathbb{N}$, are also iid random variables, Theorem 4.1 conditions are met and we can conclude that

$$\begin{aligned}\lim_{n\to\infty} \frac{g_n(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n)}{n} &= \lim_{n\to\infty} -B_n(\zeta, \delta) \\ &= -b(\zeta, \delta), \ a.s.,\end{aligned}$$

where $b(\zeta, \delta)$ is a constant depending on parameters (ζ, δ). ∎

Lemma 4.4. *The limit function $b\colon \mathbb{R}^2 \to \mathbb{R}$ is convex in $\mathbb{R}^2$.*

Proof. We know from Lemma 3.10 that B_n is convex for all $n \in \mathbb{N}$. For every $(\zeta_1, \delta_1), (\zeta_2, \delta_2) \in \mathbb{R}^2$ and $\lambda \in (0, 1)$

$$B_n(\lambda\zeta_1 + (1-\lambda)\zeta_2, \lambda\delta_1 + (1-\lambda)\delta_2) \leq \lambda B_n(\zeta_1, \delta_1) + (1-\lambda)B_n(\zeta_2, \delta_2).$$

Letting $n \to \infty$ and using Lemma 4.3, we get

$$\begin{aligned}b(\lambda\zeta_1 + (1-\lambda)\zeta_2, \lambda\delta_1 + (1-\lambda)\delta_2) &= \lim_{n\to\infty} B_n(\lambda\zeta_1 + (1-\lambda)\zeta_2, \lambda\delta_1 + (1-\lambda)\delta_2) \\ &\leq \lim_{n\to\infty} (\lambda B_n(\zeta_1, \delta_1) + (1-\lambda)B_n(\zeta_2, \delta_2) \\ &= \lambda b(\zeta_1, \delta_1) + (1-\lambda)b(\zeta_2, \delta_2).\end{aligned}$$

Hence the limit function b is convex in $\mathbb{R}^2$. ∎

Lemma 4.5. *Following holds for the limit function $b\colon \mathbb{R}^2 \to \mathbb{R}$:*

$$\mathbf{P}\left(\bigcup_{(\zeta,\delta)\in\mathbb{R}^2} \left\{\omega\colon \lim_{n\to\infty} B_n(\zeta, \delta)(\omega) = b(\zeta, \delta)\right\}\right) = 1 \tag{4.2}$$

Proof. We leave out the argument ω to simplify notation. We know from Lemma 4.3, that for each $(\zeta, \delta) \in \mathbb{R}^2$

$$\mathbf{P}\left(\left\{\lim_{n\to\infty} B_n(\zeta, \delta) = b(\zeta, \delta)\right\}\right) = 1.$$

Considering a dense countable subset $\mathbb{Q} \times \mathbb{Q} \subset \mathbb{R}^2$ and using countable additivity of probability, we get

$$\begin{aligned}\mathbf{P}\left(\bigcup_{(\zeta,\delta)\in\mathbb{Q}\times\mathbb{Q}} \left\{\lim_{n\to\infty} B_n(\zeta, \delta) \neq b(\zeta, \delta)\right\}\right) &= \sum_{(\zeta,\delta)\in\mathbb{Q}\times\mathbb{Q}} \mathbf{P}\left(\left\{\lim_{n\to\infty} B_n(\zeta, \delta) \neq b(\zeta, \delta)\right\}\right) \\ &= 0.\end{aligned} \tag{4.3}$$

Function b is continuous, since it is convex, which is stated by Lemma 4.4. Also, $\mathbb{Q} \times \mathbb{Q}$ is dense in $\mathbb{R}^2$ and using (4.3), we get (4.2). ∎

Apart from the convergence of B_n, we are also interested in the behaviour of the proportions, if $n \to \infty$. In the following section we look into the properties of asymptotic proportions.

4.2 Differentiability of b

Lemma 4.6. (see, e.g., [LMT] page 6). *Functions' $\delta \mapsto B_n(\delta)$, $n \in \mathbb{N}$, convergence $B_n \to b$ is uniform in $\mathbb{R}$.*

Uniform convergence of $\delta \mapsto B_n(\delta)$ in $\mathbb{R}$ does not guarantee the convergence of partial derivatives. Uniform convergence of convex functions only implies the convergence of one-sided partial derivatives at values, where b is differentiable. There are examples of uniformly converging sequences of functions with diverging sequence of one-sided derivatives. The following example demonstrates this.

Example 4.7. Let $f_n \colon \mathbb{R} \to \mathbb{R}$, $n \in \mathbb{N}$ be a sequence of convex functions

$$f_n(x) = \left| x - \frac{(-1)^n}{n} \right|.$$

Using reverse triangle inequality, we get

$$\left| \underbrace{\left| x - \frac{(-1)^n}{n} \right|}_{f_n(x)} - \underbrace{|x|}_{f(x)} \right| \leq \left| x - \frac{(-1)^n}{n} - x \right| = \frac{1}{n},$$

hence $f_n \to f$ converges uniformly in $\mathbb{R}$. One sided derivatives do not converge at $x = 0$, since

$$f_n'(0_+) = (-1)^n,$$
$$f_n'(0_-) = (-1)^n.$$

Next, we will consider B_n as a two-variable function again. Instead of convergence of one-sided derivatives for b, we have the following inequalities.

Lemma 4.8. *Following inequalities hold:*[1]

$$\frac{\partial b(\zeta, \delta_-)}{\partial \delta} \leq \liminf_n \underline{q}_n(\zeta, \delta) \leq \limsup_n \overline{q}_n(\zeta, \delta) \leq \frac{\partial b(\zeta, \delta_+)}{\partial \delta}, \qquad a.e.$$
$$\frac{\partial b(\zeta_-, \delta)}{\partial \zeta} \leq \liminf_n \underline{p}_n(\zeta, \delta) \leq \limsup_n \overline{p}_n(\zeta, \delta) \leq \frac{\partial b(\zeta_+, \delta)}{\partial \zeta}, \qquad a.e.$$

Therefore if b has partial derivatives at (ζ, δ), then following proportions converge, as $n \to \infty$:

$$\underline{q}_n(\zeta, \delta) \to \frac{\partial b(\zeta, \delta)}{\partial \delta} \qquad \text{and} \qquad \overline{q}_n(\zeta, \delta) \to \frac{\partial b(\zeta, \delta)}{\partial \delta}, \qquad a.s.$$
$$\underline{p}_n(\zeta, \delta) \to \frac{\partial b(\zeta, \delta)}{\partial \zeta} \qquad \text{and} \qquad \overline{p}_n(\zeta, \delta) \to \frac{\partial b(\zeta, \delta)}{\partial \zeta}, \qquad a.s.$$

[1] First line of inequalities is analogously proven in [LMT] (3.2).

Proof. For all $(\zeta, \delta) \in \mathbb{R}^2$ we have

$$\begin{aligned}
\frac{\partial b(\zeta, \delta_-)}{\partial \delta} &= \lim_{s \to 0-} \frac{b(\zeta, \delta + s) - b(\zeta, \delta)}{s} \\
&= \lim_{s \to 0-} \lim_{n \to \infty} \frac{B_n(\zeta, \delta + s) - B_n(\zeta, \delta)}{s} \\
&\leq \liminf_{n \to \infty} \lim_{s \to 0-} \frac{B_n(\zeta, \delta + s) - B_n(\zeta, \delta)}{s} \qquad (4.4) \\
&= \liminf_{n \to \infty} \underline{q}_n(\zeta, \delta) \\
&\leq \limsup_{n \to \infty} \overline{q}_n(\zeta, \delta) \\
&= \limsup_{n \to \infty} \lim_{s \to 0+} \frac{B_n(\zeta, \delta + s) - B_n(\zeta, \delta)}{s} \\
&\leq \lim_{s \to 0+} \limsup_{n \to \infty} \frac{B_n(\zeta, \delta + s) - B_n(\zeta, \delta)}{s} \qquad (4.5) \\
&= \frac{\partial b(\zeta, \delta_+)}{\partial \delta}.
\end{aligned}$$

Equalities (4.5) and (4.4) hold according to [Roc] Theorem 24.5. Analogously, for all $(\zeta, \delta) \in \mathbb{R}$, we also get

$$\begin{aligned}
\frac{\partial b(\zeta_-, \delta)}{\partial \zeta} &= \lim_{s \to 0-} \frac{b(\zeta + s, \delta) - b(\zeta, \delta)}{s} \\
&= \lim_{s \to 0-} \lim_{n \to \infty} \frac{B_n(\zeta + s, \delta) - B_n(\zeta, \delta)}{s} \\
&\leq \liminf_{n \to \infty} \lim_{s \to 0-} \frac{B_n(\zeta + s, \delta) - B_n(\zeta, \delta)}{s} \\
&= \liminf_{n \to \infty} \underline{p}_n(\zeta, \delta) \\
&\leq \limsup_{n \to \infty} \overline{p}_n(\zeta, \delta) \\
&= \limsup_{n \to \infty} \lim_{s \to 0+} \frac{B_n(\zeta + s, \delta) - B_n(\zeta, \delta)}{s} \\
&\leq \lim_{s \to 0+} \limsup_{n \to \infty} \frac{B_n(\zeta + s, \delta) - B_n(\zeta, \delta)}{s} \\
&= \frac{\partial b(\zeta, \delta_+)}{\partial \zeta}.
\end{aligned}$$

■

Although we cannot prove the convergence of proportions, Lebesque measure of the set, where the function b is not differentiable, is actually zero. Next lemma clarifies this notion.

Lemma 4.9. *Limit function* $b\colon \mathbb{R}^2 \to \mathbb{R}$ *is differentiable almost everywhere in* $\mathbb{R}^2$.

Proof. Lemma 4.4 states that b is convex in $\mathbb{R}^2$. We know (see, e.g., [Roc], Theorem 25.5) that convex function is almost everywhere differentiable. ∎

4.3 Simulations

For the following simulations we have used a binary alphabet $\mathbb{A} = \{a, b\}$ and generated uniformly random sequences with length n. Considering the extensive memory usage of Needleman-Wunsch algorithm in implementation with Python, we have restricted our simulations to $n = 10,\ 100,\ 500,\ 1000$. First, we fixed have $\zeta^* = 0.5$ to plot the graph of $\delta \mapsto q_n(\zeta^*, \delta)$.

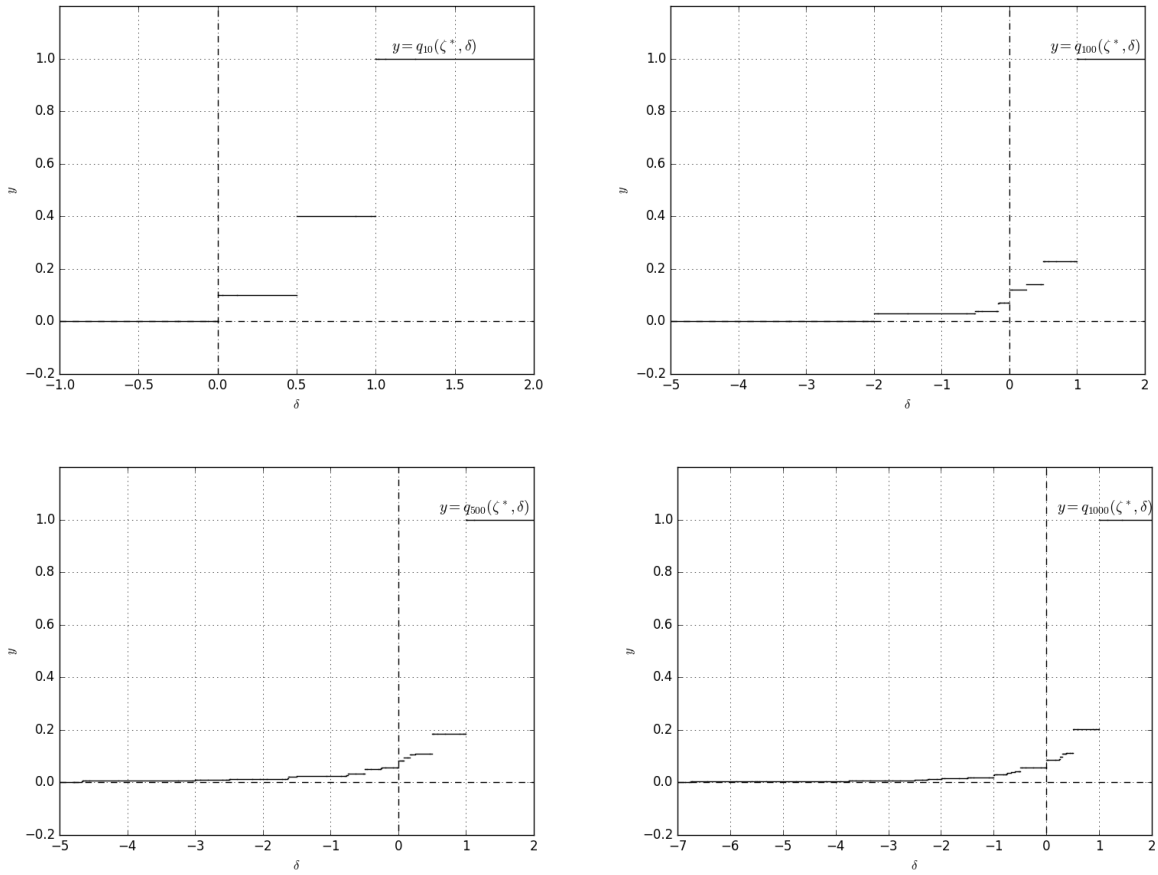

Figure 4.10: Graphs of $\delta \mapsto q_n(\zeta^*, \delta)$, where $n = 10,\ 100,\ 500,\ 1000$

Figure 4.10 illustrates the asymptotic behaviour of gaps. We can see that the difference between $n = 500$ and $n = 1000$ is almost negligible and differences between $\overline{q}_n$ and $\underline{q}_n$ decreases rapidly. There are still 2 values at which the "jump" occurs in the graph. At first the gap price exceeds the score of the mismatch, substituting each mismatch in the optimal alignment with a gap. The second "jump" occurs when the optimal alignment consists only of gaps.

Next, we have fixed $\delta^* = -2$ to plot the graph of $\zeta \mapsto q_n(\zeta, \delta^*)$.

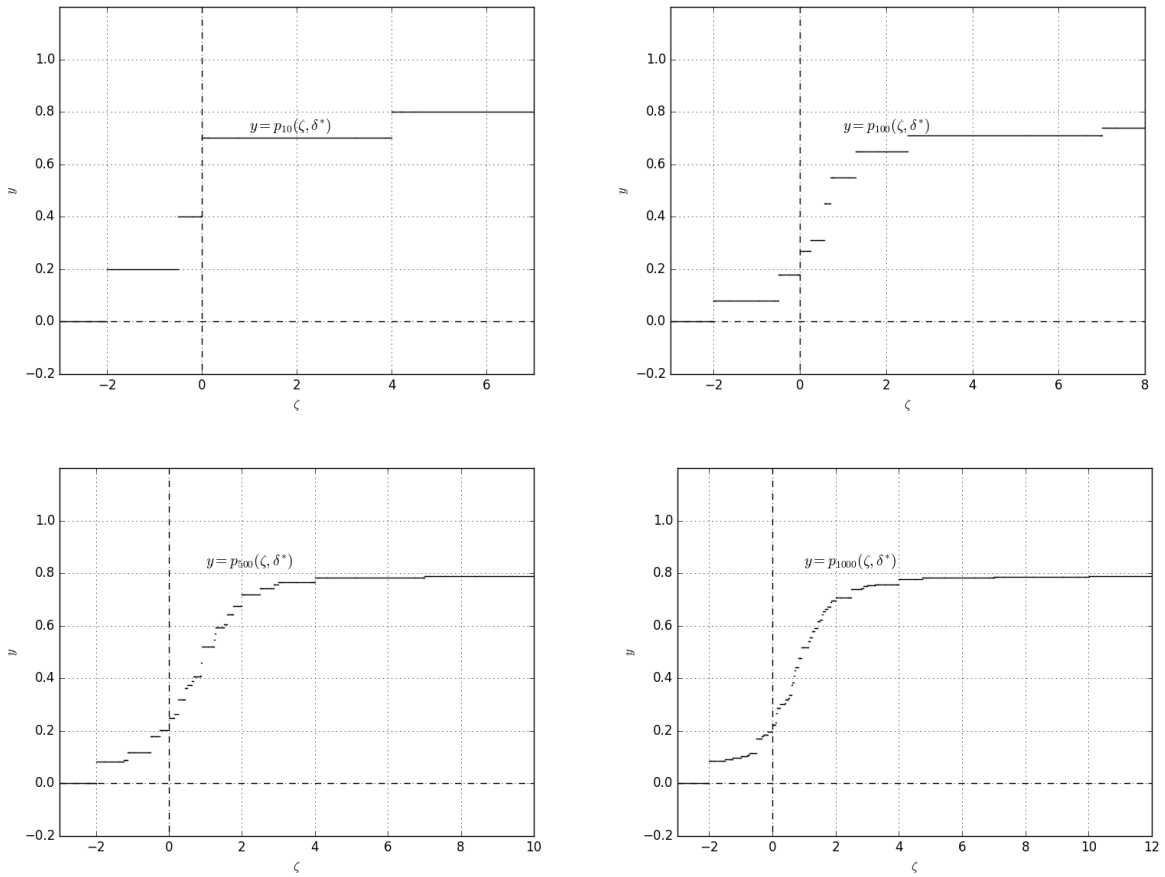

Figure 4.11: Graphs of $\zeta \mapsto p_n(\zeta, \delta^*)$, where $n = 10,\ 100,\ 500,\ 1000$

Figure 4.11 illustrates the asymptotic behaviour of mismatches. We can see that the difference between $n = 500$ and $n = 1000$ is almost negligible and difference between $\overline{p}_n$ and $\underline{p}_n$ decreases rapidly. There is still a single value at which the "jump" occurs in the graph. It happens when the score of mismatched letters exceeds the gap price, substituting gaps in the optimal alignment with possible mismatched letters.

Let us fix $\zeta^{**} = -1.5$ and for the next figure, we generated random sequences, from uniform distribution, of length $n = 1000$ to illustrate the asymptotic relation between $\underline{p}_n$, $\overline{p}_n$.

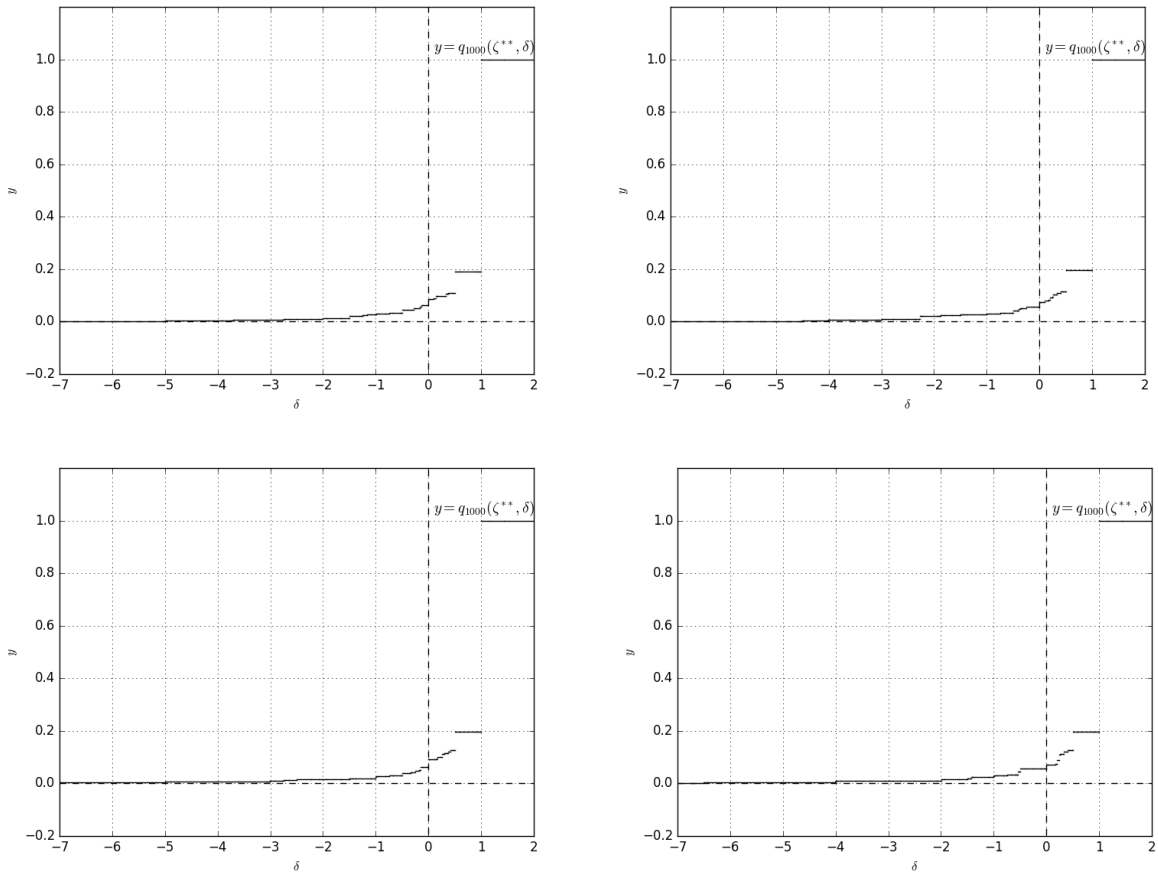

Figure 4.12: Graphs of $\delta \mapsto q_{1000}(\zeta^{**}, \delta)$, for 4 randomly generated sequences

Let us fix $\delta^{**} = -1$ to plot the graph of a function $\zeta \mapsto p_{1000}(\zeta, \delta^{**})$.

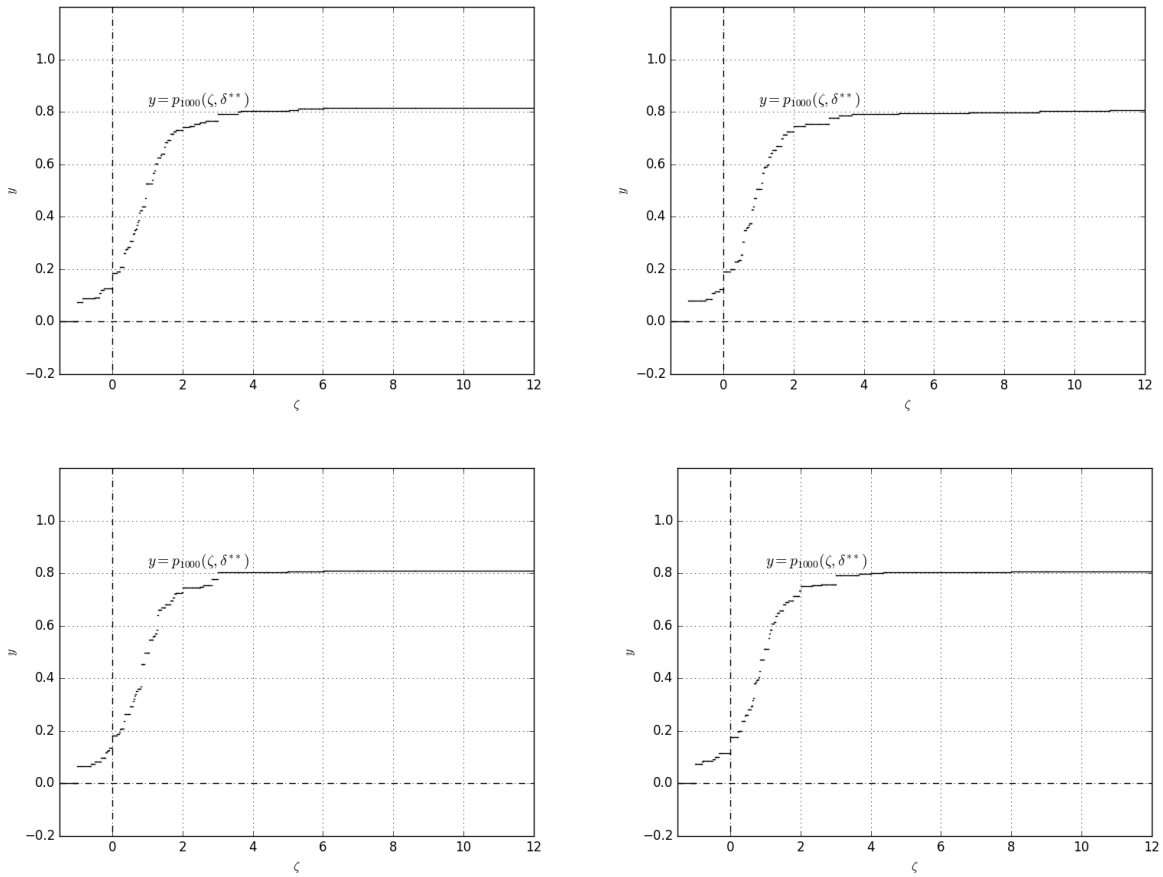

Figure 4.13: Graphs of $\zeta \mapsto p_{1000}(\zeta, \delta^{**})$, for 4 randomly generated sequences

Figure 4.12 and Figure 4.13 display the negligible change when generating different random sequences. This helps to confirm the theory that we developed in this thesis.

Chapter 5

Large Deviations

In this chapter, we prove the bound on the score of the optimal alignment, when changing a single letter. Given $(\zeta, \delta) \in \mathbb{R}^2$, we derive large deviations inequality for one-sided partial derivatives of B_n.

Alphabet is still fixed as $\mathbb{A} = \{a, b\}$ and

$$S(a,b) = S(b,a) = \zeta, \quad S(a,a) = S(b,b) = 1.$$

5.1 Bound on the Score, When Changing a Single Letter

First, we need to introduce the maximum change of optimal alignment score, when changing a single letter in either sequence x or y. We denote it with $\Delta_{(\zeta,\delta)}$. Such value is defined, in order to use it when applying McDiarmid's inequality. We shall abbreviate the effect of changing a single letter by

$$\Delta_{(\zeta,\delta)} := \sup_{\substack{x,y \in A^n \\ x_i^* \in A \vee y_i^* \in A}} \left| B_n(\zeta,\delta) - B_n^*(\zeta,\delta) \right|,$$

where the supremum is taken over all possible sequences x and y, while varying a single letter (denoted by x_i^* or y_i^*) at any single position $i \in \{1, 2, \dots n\}$ in x or y. $B_n(\zeta,\delta)$ denotes the unchanged sequences score and $B_n^*(\zeta,\delta)$ the score of sequences with one changed letter (either with letter x_i^* or y_i^*).

Lemma 5.1. *Following holds for all $(\zeta,\delta) \in \mathbb{R}^2$:*

$$\Delta_{(\zeta,\delta)} \leq |1 - \zeta|. \tag{5.1}$$

Proof. Let us fix sequences $x, y \in A^n$, parameters $(\zeta,\delta) \in \mathbb{R}^2$ and a position in sequences $i \in \{0, 1, 2, \dots, n\}$. Changing the letter at i-th position can either increase or decrease the optimal alignment score. Let us consider both cases separately.

Let us assume the optimal alignment score decreases. If one of the optimal alignments has an indel at i-th position, changing the letter at i-th position can always leave the indel intact. Hence the optimal alignment score cannot decrease. If none of the optimal alignments have an indel at i-th position, the change in score is

$$\begin{aligned}\Delta_{(\zeta,\delta)}^{\text{decrease}} &\leq \max_{x,y,z} |S(x,y) - S(y,z)| \\ &= |1-\zeta|. \end{aligned} \tag{5.2}$$

Let us assume the optimal alignment score increases. In that case, we can consider the new optimal alignment score instead. We know that the decrease of the new optimal alignment score is

$$\begin{aligned}\Delta_{(\zeta,\delta)}^{\text{increase}} &= \Delta_{(\zeta,\delta)}^{\text{new score decrease}} \\ &\leq \max_{x,y,z} |S(x,y) - S(y,z)| \qquad (5.3) \\ &= |1-\zeta|. \qquad (5.4)\end{aligned}$$

Inequality (5.3) is the result of the previous case, where the optimal alignment score decreased.

Results (5.2) and (5.4) give us (5.1). ∎

We use McDiarmid's inequality to derive large deviations inequality for the proportions.

Theorem 5.2 (McDiarmid's inequality)**.** (see, e.g., [DGG], page 136)**.**
Let $Z_1, Z_2, \ldots, Z_{2m}$ be iid random variables taking values in a set A, and assume $f\colon A^{2n} \to \mathbb{R}$ satisfies

$$\sup_{x_1,x_2,\ldots,x_{2n}\in A,\ x_i^*\in A} |f(x_1,x_2,\ldots,x_{2n}) - f(x_1,x_2,\ldots,x_{i-1},x_i^*,x_{i+1},\ldots,x_{2n})| \leq K,$$

for $1 \leq i \leq 2n$, then for all $\varepsilon > 0$

$$\mathbf{P}(|f(Z_1,Z_2,\ldots,Z_{2n}) - \mathbf{E}f(Z_1,Z_2,\ldots,Z_{2n})| > \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{nK}\right).$$

In our case we apply Theorem 5.2, with Lemma 5.1, as $Z_1 = X_1,\ Z_2 = X_2, \ldots, Z_n = X_n, Z_{n+1} = Y_1,\ \ldots, Z_{2n} = Y_n$ and f is in the role of B_n. Hence, for every fixed $(\zeta,\delta) \in \mathbb{R}^2$, $\zeta \neq 1$ and $\varepsilon > 0$

$$\mathbf{P}(B_n(\zeta,\delta) - \mathbf{E}B_n(\zeta,\delta) \leq -\varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{|1-\zeta|^2}\right), \tag{5.5}$$

$$\mathbf{P}(B_n(\zeta,\delta) - \mathbf{E}B_n(\zeta,\delta) \geq \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{|1-\zeta|^2}\right). \tag{5.6}$$

For every $(\zeta,\delta) \in \mathbb{R}^2$ the function B_n at (ζ,δ) is always bounded. Therefore we can use dominated convergence theorem (see, e.g., [Bil], Theorem 16.4), and

$$\lim_{n\to\infty} \mathbf{E}B_n(\zeta,\delta) = b(\zeta,\delta),$$

for all $(\zeta,\delta) \in \mathbb{R}^2$.

5.2 Large Deviations Inequality

Theorem 5.3. [1] *Let $(\zeta, \delta) \in \mathbb{R}^2$, $\zeta \neq 1$. For every $\varepsilon > 0$ there exists $n(\varepsilon) \in \mathbb{N}$ and $c(\varepsilon, \zeta, \delta) \geq 0$, such that such that*

$$\mathbf{P}\left(\frac{\partial b(\zeta_-, \delta)}{\partial \zeta} - \varepsilon \leq \underline{p}_n(\zeta, \delta) \leq \overline{p}_n(\zeta, \delta) \leq \frac{\partial b(\zeta_+, \delta)}{\partial \zeta} + \varepsilon\right) \geq 1 - 4\exp\left(-\frac{n\varepsilon^2 c(\varepsilon, \zeta, \delta)}{16}\right), \tag{5.7}$$

Proof. Given $(\zeta, \delta) \in \mathbb{R}^2$, $\zeta \neq 1$ and $\varepsilon > 0$, the first step is to derive bounds on

$$\mathbf{P}\left(\frac{\partial B_n(\zeta_+, \delta)}{\partial \zeta} - \frac{\partial b(\zeta_+, \delta)}{\partial \zeta} > \varepsilon\right).$$

For $s_1(\varepsilon) > 0,\ s_2(\varepsilon) > 0$ and any function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$, let us denote

$$\Delta\varphi_+ := \varphi(\zeta + s_1(\varepsilon), \delta) - \varphi(\zeta, \delta),$$

$$\Delta\varphi_- := \varphi(\zeta - s_2(\varepsilon), \delta) - \varphi(\zeta, \delta).$$

Since

$$\lim_{s \to 0+} \frac{b(\zeta + s, \delta) - b(\zeta, \delta)}{s} = \frac{\partial b(\zeta_+, \delta)}{\partial \zeta},$$

we can choose $s_1(\varepsilon) \in (0, 1)$ such that

$$\left|\frac{\Delta b_+}{s_1(\varepsilon)} - \frac{\partial b(\zeta_+, \delta)}{\partial \zeta}\right| \leq \frac{\varepsilon}{4}. \tag{5.8}$$

Next, we will fix $N_1 \in \mathbb{N}$ (also depending on ε), such that for all $n \geq N_1$

$$|\Delta \mathbf{E} B_{n+} - \Delta b_+| \leq s_1(\varepsilon)\frac{\varepsilon}{4}.$$

Thus for those $s_1(\varepsilon)$ and n chosen, we have

$$\begin{aligned}\frac{\Delta B_{n+}}{s_1(\varepsilon)} - \frac{\partial b(\zeta_+, \delta)}{\partial \zeta} &= \left(\frac{\Delta B_{n+}}{s_1(\varepsilon)} - \frac{\Delta \mathbf{E} B_{n+}}{s_1(\varepsilon)}\right) \\ &\quad + \left(\frac{\Delta \mathbf{E} B_{n+}}{s_1(\varepsilon)} - \frac{\Delta b_+}{s_1(\varepsilon)}\right) + \left(\frac{\Delta b_+}{s_1(\varepsilon)} - \frac{\partial b(\zeta_+, \delta)}{\partial \zeta}\right) \\ &\leq \left(\frac{\Delta B_{n+}}{s_1(\varepsilon)} - \frac{\Delta \mathbf{E} B_{n+}}{s_1(\varepsilon)}\right) + \frac{\varepsilon}{2}. \end{aligned} \tag{5.9}$$

[1] This theorem is analogous to [LMT] Theorem 4.1

From (5.5) and (5.6) we get

$$\begin{aligned}\mathbf{P}\left(\frac{\Delta B_{n+}}{s_1(\varepsilon)} - \frac{\Delta \mathbf{E}B_{n+}}{s_1(\varepsilon)} \geq \frac{\varepsilon}{2}\right) &= \mathbf{P}\Big((B_n(\zeta + s_1(\varepsilon), \delta) - B_n(\zeta, \delta)) \\ &\quad - (\mathbf{E}B_n(\zeta + s_1(\varepsilon), \delta) - \mathbf{E}B_n(\zeta, \delta)) \geq s_1(\varepsilon)\frac{\varepsilon}{2}\Big) \\ &\leq \mathbf{P}\left(B_n(\zeta, \delta) - \mathbf{E}B_n(\zeta, \delta) \leq -s_1(\varepsilon)\frac{\varepsilon}{4}\right) \\ &\quad + \mathbf{P}\left(B_n(\zeta + s_1(\varepsilon), \delta) - \mathbf{E}B_n(\zeta + s_1(\varepsilon), \delta) \geq s_1(\varepsilon)\frac{\varepsilon}{4}\right) \qquad (5.10) \\ &\leq 2\exp\left(-\frac{n s_1^2(\varepsilon)\varepsilon^2}{16 A_+^2(\zeta, \varepsilon)}\right), \qquad (5.11)\end{aligned}$$

where $A_+(\zeta, \varepsilon) := \max\{|1-\zeta|, |1-(\zeta + s_1(\varepsilon))|\}$. For explanation, value $\zeta + s_1(\zeta)$ arose at (5.8) and we considered it at (5.10), when applying the McDiarmid's inequality. Since B_n is convex, we get

$$\frac{\partial B_n(\zeta_+, \delta)}{\partial \zeta} \leq \frac{\Delta B_{n+}}{s_1(\varepsilon)},$$

for ε and $s_1(\varepsilon)$ chosen as in (5.9), (5.11) and

$$\begin{aligned}\mathbf{P}\left(\frac{\partial B_n(\zeta_+, \delta)}{\partial \zeta} - \frac{\partial b(\zeta_+, \delta)}{\partial \zeta} \geq \varepsilon\right) &\leq \mathbf{P}\left(\frac{\Delta B_{n+}}{s_1(\varepsilon)} - \frac{\partial b(\zeta_+, \delta)}{\partial \zeta} \geq \varepsilon\right) \\ &\leq \mathbf{P}\left(\frac{\Delta B_{n+}}{s_1(\varepsilon)} - \frac{\Delta \mathbf{E}B_{n+}}{s_1(\varepsilon)} + \frac{\varepsilon}{2} \geq \varepsilon\right) \\ &= \mathbf{P}\left(\frac{\Delta B_{n+}}{s_1(\varepsilon)} - \frac{\Delta \mathbf{E}B_{n+}}{s_1(\varepsilon)} \geq \frac{\varepsilon}{2}\right) \\ &\leq 2\exp\left(-\frac{n s_1^2(\varepsilon)\varepsilon^2}{16 A_+^2(\zeta, \varepsilon)}\right). \qquad (5.12)\end{aligned}$$

Next step is to derive bounds on

$$\mathbf{P}\left(\frac{\partial B_n(\zeta_-, \delta)}{\partial \zeta} - \frac{\partial b(\zeta_-, \delta)}{\partial \zeta} < -\varepsilon\right).$$

Since

$$\lim_{s \to 0+} \frac{b(\zeta - s, \delta) - b(\zeta, \delta)}{-s} = \frac{\partial b(\zeta_-, \delta)}{\partial \zeta},$$

we can choose $s_2(\varepsilon) \in (0, 1)$, such that

$$\left|\frac{\partial b(\zeta_-, \delta)}{\partial \zeta} - \frac{\Delta b_-}{-s_2(\varepsilon)}\right| \leq \frac{\varepsilon}{4}. \qquad (5.13)$$

We will fix $N_2 \in \mathbb{N}$ (also depending on ε), such that for all $n \geq N_2$

$$|\Delta \mathbf{E}B_{n-} - \Delta b_-| \leq s_2(\varepsilon)\frac{\varepsilon}{4}.$$

Let us denote $N := \max\{N_1, N_2\}$. Thus, for those $s_2(\varepsilon)$ and $n \geq N$, we get

$$\begin{aligned}\frac{\Delta B_{n-}}{-s_2(\varepsilon)} - \frac{\partial b(\zeta_-, \delta)}{\partial \zeta} &= \left(\frac{\Delta B_{n-}}{-s_2(\varepsilon)} - \frac{\Delta \mathbf{E} B_{n-}}{-s_2(\varepsilon)}\right) \\ &\quad - \left(\frac{\Delta \mathbf{E} B_{n-}}{s_2(\varepsilon)} - \frac{\Delta b_-}{s_2(\varepsilon)}\right) - \left(\frac{\partial b(\zeta_-, \delta)}{\partial \zeta} - \frac{\Delta b_-}{-s_2(\varepsilon)}\right) \\ &\geq \left(\frac{\Delta B_{n-}}{-s_2(\varepsilon)} + \frac{\Delta \mathbf{E} B_{n-}}{-s_2(\varepsilon)}\right) - \frac{\varepsilon}{2}. \end{aligned} \tag{5.14}$$

From (5.5) and (5.6) we get

$$\begin{aligned}\mathbf{P}\left(\frac{\Delta B_{n-}}{-s_2(\varepsilon)} - \frac{\Delta \mathbf{E} B_{n-}}{-s_2(\varepsilon)} \leq -\frac{\varepsilon}{2}\right) &= \mathbf{P}\Big((B_n(\zeta - s_2(\varepsilon), \delta) - B_n(\zeta, \delta)) \\ &\quad - (\mathbf{E} B_n(\zeta - s_2(\varepsilon), \delta) - \mathbf{E} B_n(\zeta, \delta)) \geq s_2(\varepsilon)\frac{\varepsilon}{2}\Big) \\ &\leq \mathbf{P}\left(B_n(\zeta, \delta) - \mathbf{E} B_n(\zeta, \delta) \leq -s_2(\varepsilon)\frac{\varepsilon}{4}\right) \\ &\quad + \mathbf{P}\left(B_n(\zeta - s_2(\varepsilon), \delta) - \mathbf{E} B_n(\zeta - s_2(\varepsilon), \delta) \geq s_2(\varepsilon)\frac{\varepsilon}{4}\right) \end{aligned} \tag{5.15}$$

$$\leq 2\exp\left(-\frac{n s_2^2(\varepsilon)\varepsilon^2}{16 A_-^2(\zeta, \varepsilon)}\right), \tag{5.16}$$

where $A_-(\zeta, \varepsilon) := \max\{|1 - \zeta|, |1 - (\zeta - s_2(\varepsilon))|\}$. For explanation, value $\zeta - s_2(\zeta)$ arose at (5.13) and we considered it at (5.15), when applying the McDiarmid's inequality. Since B_n is convex, we get

$$\frac{\partial B_n(\zeta_-, \delta)}{\partial \zeta} \geq \frac{\Delta B_{n-}}{-s_2(\varepsilon)},$$

for ε and $s_2(\varepsilon)$ chosen as in (5.14), (5.16) and

$$\begin{aligned}\mathbf{P}\left(\frac{\partial B_n(\zeta_-, \delta)}{\partial \zeta} - \frac{\partial b(\zeta_-, \delta)}{\partial \zeta} \leq -\varepsilon\right) &\leq \mathbf{P}\left(\frac{\Delta B_{n-}}{-s_2(\varepsilon)} - \frac{\partial b(\zeta_-, \delta)}{\partial \zeta} \leq -\varepsilon\right) \\ &\leq \mathbf{P}\left(\frac{\Delta B_{n-}}{-s_2(\varepsilon)} - \frac{\Delta \mathbf{E} B_{n-}}{-s_2(\varepsilon)} - \frac{\varepsilon}{2} \leq -\varepsilon\right) \\ &= \mathbf{P}\left(\frac{\Delta B_{n-}}{-s_2(\varepsilon)} - \frac{\Delta \mathbf{E} B_{n-}}{-s_2(\varepsilon)} \leq -\frac{\varepsilon}{2}\right) \\ &\leq 2\exp\left(-\frac{n s_2^2(\varepsilon)\varepsilon^2}{16 A_-^2(\zeta, \varepsilon)}\right). \end{aligned} \tag{5.17}$$

From (5.12) and (5.17) we get

$$\mathbf{P}\left(\overline{p}_n(\zeta, \delta) > \frac{\partial b(\zeta_-, \delta)}{\partial \zeta} + \varepsilon\right) \geq 2\exp\left(-\frac{n s_1^2(\varepsilon)\varepsilon^2}{16 A_+^2(\zeta, \varepsilon)}\right), \tag{5.18}$$

$$\mathbf{P}\left(\underline{p}_n(\zeta,\delta) < \frac{\partial b(\zeta_-,\delta)}{\partial\zeta} - \varepsilon\right) \geq 2\exp\left(-\frac{ns_2^2(\varepsilon)\varepsilon^2}{16A_-^2(\zeta,\varepsilon)}\right). \tag{5.19}$$

Combining (5.19), (5.18) and using the probability of a complementary event, gives us (5.7). ∎

Remark 5.4. *More precisely,*

$$c(\varepsilon,\zeta,\delta) = \frac{(s(\varepsilon))^2}{A^2(\zeta,\delta)},$$

where $s(\varepsilon) = \min\{s_1(\varepsilon), s_2(\varepsilon)\} > 0$ *and* $A(\zeta,\varepsilon) := \max\{A_+(\zeta,\delta), A_-(\zeta,\delta)\} > 0$.

Chapter 6

Function $(\zeta, \delta) \mapsto B_n(\zeta, \delta)$ as $\delta_1(\zeta, \delta) \mapsto B_n(\zeta', \delta_1(\zeta, \delta))$

In the following chapter we introduce the idea of considering the two-variable function B_n as a single variable function. We fix the parameter ζ and derive some two-variable case results from the one-variable case.

Let us denote a function

$$\delta_1(\zeta, \delta) := \frac{\delta(1-\zeta') + \zeta' - \zeta}{1-\zeta}. \tag{6.1}$$

If $\zeta < 1$, then for all $\zeta' < 1$

$$\begin{aligned}
B_n(\zeta, \delta) &= \max_{(p,q)\in\overline{\mathcal{O}}'_n} (1 + p(\zeta - 1) + q(\delta - 1)) \\
&= \max_{(p,q)\in\overline{\mathcal{O}}'_n} \frac{(1-\zeta') + p(1-\zeta)(\zeta'-1) + q(\delta-1)(1-\zeta')}{1-\zeta'} \\
&= \max_{(p,q)\in\overline{\mathcal{O}}'_n} \frac{\left(1 + p(\zeta'-1) + \frac{q(\delta-1)(1-\zeta')}{1-\zeta}\right)(1-\zeta) + \zeta - \zeta'}{1-\zeta'} \\
&= \max_{(p,q)\in\overline{\mathcal{O}}'_n} \frac{\left(1 + p(\zeta'-1) + q\left(\frac{\delta(1-\zeta')+\zeta'-\zeta}{1-\zeta} - 1\right)\right)(1-\zeta) + \zeta - \zeta'}{1-\zeta'} \\
&= \frac{B_n(\zeta', \delta_1(\zeta, \delta))(1-\zeta) + \zeta - \zeta'}{1-\zeta'}
\end{aligned} \tag{6.2}$$

Equality (6.2) allows us to reduce the two-variable function B_n to a one-variable function and $\mathcal{O}_n(\zeta, \delta) = \mathcal{O}_n(\zeta', \delta_1(\zeta, \delta))$. Hence, we can find all optimal alignments through varying only the gap price. Let us fix $\zeta' = 0$ and restrict $\zeta < 1$ throughout this chapter. Using (6.2), the two-variable function B_n is

$$B_n(\zeta, \delta) = B_n(0, \delta_1(\zeta, \delta))(1-\zeta) + \zeta, \tag{6.3}$$

for $\delta \in \mathbb{R}$ and (6.1) becomes

$$\delta_1(\zeta, \delta) = \frac{\delta - \zeta}{1 - \zeta}.$$

Lemma 6.1. *Function $\delta \mapsto B_n(\zeta, \delta)$ is differentiable at $\delta \in \mathbb{R}$ iff $\delta \mapsto B_n(0, \delta)$ is differentiable at $\delta_1(\zeta, \delta)$ and*

$$\frac{\partial B_n(\zeta, \delta)}{\partial \delta} = \frac{dB_n(0, \delta_1(\zeta, \delta))}{d\delta_1} = q_n(0, \delta_1(\zeta, \delta)).$$

Proof. Let $\delta \mapsto B_n(\zeta, \delta)$ be differentiable at $\delta \in \mathbb{R}$. Using (6.3), we get

$$\begin{aligned}
\frac{\partial B_n(\zeta, \delta)}{\partial \delta} &= \frac{\partial}{\partial \delta} \left(B_n(0, \delta_1(\zeta, \delta))(1 - \zeta) + \zeta \right) \\
&= (1 - \zeta) \frac{d}{d\delta_1} B_n(0, \delta_1(\zeta, \delta)) \frac{\partial \delta_1(\zeta, \delta)}{\partial \delta} \\
&= (1 - \zeta) \frac{d}{d\delta_1} B_n(0, \delta_1(\zeta, \delta)) \frac{1}{1 - \zeta} \\
&= q_n(0, \delta_1(\zeta, \delta))
\end{aligned}$$

and $\delta \mapsto B_n(0, \delta)$ is differentiable at $\delta_1(\zeta, \delta)$. Let $\delta \mapsto B_n(0, \delta)$ be differentiable at $\delta_1(\zeta, \delta)$. Differentiability of $\delta \mapsto B_n(\zeta, \delta)$ at δ follows from (6.3). ∎

Lemma 6.2. *Function $\zeta \mapsto B_n(\zeta, \delta)$ is differentiable at $\zeta < 1$ iff $\delta \mapsto B_n(0, \delta)$ is differentiable at $\delta_1(\zeta, \delta)$ and*

$$\frac{\partial B_n(\zeta, \delta)}{\partial \zeta} = p_n(0, \delta_1(\zeta, \delta)).$$

Proof. Let $\zeta \mapsto B_n(\zeta, \delta)$ be differentiable at $\zeta < 1$. Using (6.3), we get

$$\begin{aligned}
\frac{\partial B_n(\zeta, \delta)}{\partial \zeta} &= \frac{\partial}{\partial \zeta} \left(B_n(0, \delta_1(\zeta, \delta))(1 - \zeta) + \zeta \right) \\
&= q_n(0, \delta_1(\zeta, \delta)) \frac{\partial \delta_1(\zeta, \delta)}{\partial \zeta} (1 - \zeta) - B_n(0, \delta_1(\zeta, \delta)) + 1 \\
&= q_n(0, \delta_1(\zeta, \delta)) \frac{-(1 - \zeta) + (\delta - \zeta)}{1 - \zeta} - B_n(0, \delta_1(\zeta, \delta)) + 1 \\
&= q_n(0, \delta_1(\zeta, \delta)) \frac{\delta - 1}{1 - \zeta} + p_n(0, \delta_1(\zeta, \delta)) \\
&\quad - q_n(0, \delta_1(\zeta, \delta)) \left(\frac{\delta - \zeta}{1 - \zeta} - 1 \right) \\
&= q_n(0, \delta_1(\zeta, \delta)) \frac{\delta - 1}{1 - \zeta} + p_n(0, \delta_1(\zeta, \delta)) - q_n(0, \delta_1(\zeta, \delta)) \frac{\delta - 1}{1 - \zeta} \\
&= p_n(0, \delta_1(\zeta, \delta)).
\end{aligned}$$

and $\delta \mapsto B_n(0, \delta)$ is differentiable at $\delta_1(\zeta, \delta)$. Let $\delta \mapsto B_n(0, \delta)$ be differentiable at $\delta_1(\zeta, \delta)$. Differentiability of $\zeta \mapsto B_n(\zeta, \delta)$ at ζ follows from (6.3). ∎

Since $\mathcal{O}_n(\zeta,\delta) = \mathcal{O}_n(0,\delta_1(\zeta,\delta))$, we get

$$q_n(0,\delta_1(\zeta,\delta)) = q_n(\zeta,\delta), \tag{6.4}$$

$$p_n(0,\delta_1(\zeta,\delta)) = p_n(\zeta,\delta). \tag{6.5}$$

If $\zeta < 1$ and $\delta < 1$, then

$$\begin{aligned} B_n(\zeta,\delta) &= 1 + \underline{p}_n(\zeta,\delta)(\zeta-1) + \overline{q}_n(\zeta,\delta)(\delta-1) && (6.6)\\ &= 1 + \overline{q}_n(\zeta,\delta)(\zeta-1) + \underline{q}_n(\zeta,\delta)(\delta-1) && (6.7)\\ &= 1 + \frac{\partial B_n(\zeta-,\delta)}{\partial\zeta}(\zeta-1) + \frac{\partial B_n(\zeta,\delta+)}{\partial\delta}(\delta-1)\\ &= 1 + \frac{\partial B_n(\zeta+,\delta)}{\partial\zeta}(\zeta-1) + \frac{\partial B_n(\zeta,\delta-)}{\partial\delta}(\delta-1). \end{aligned}$$

Equations (6.6) and (6.7) are due to $(\zeta-1) < 0$, $(\delta-1) < 0$ and $p(\zeta-1)+q(\delta-1)$ being uniquely determined for all $(p,q) \in \mathcal{O}_n(\zeta,\delta)$. Hence, if B_n is differentiable at $\delta_1(\zeta,\delta)$, it has both partial derivatives and

$$(1-\zeta)\frac{\partial B_n(\zeta,\delta)}{\partial\zeta} + (1-\delta)\frac{\partial B_n(\zeta,\delta)}{\partial\delta} = 1 - B_n(\zeta,\delta). \tag{6.8}$$

The theory can be developed further by considering one-sided derivatives or the limit function b, in the context of varying a single variable $\delta_1(\zeta,\delta)$.

Appendix

Needleman-Wunsh algorithm

The following algorithm is adjusted for alphabet $\mathbb{A} = \{a, b\}$, fixed sequences X, Y and the gap price δ, as a variable. For calculating B_n, the author has opted for programming language Python.

Listing I: Needleman-Wunsh

```
#Fixed sequences
X="abaab"
Y="baabb"

#Score parameters
zeta=0
delta=-1

def trace(delta,zeta,t,r,str1,str2,x,y,s1='',s2=''):
if x > 0 or y > 0:
c = t[y][x]
u = c == (t[y - 1][x] + delta/2)
l = c == (t[y][x - 1] + delta/2)
v = str1[x - 1] == str2[y - 1]
ul = c == (t[y - 1][x - 1] + 1 if v else
t[y - 1][x - 1] + zeta)
if ul:
trace(delta, zeta, t, r, str1, str2,
x - 1, y - 1, str1[x - 1] + s1, str2[y - 1] + s2)
elif l:
trace(delta, zeta, t, r, str1, str2,
x - 1, y, str1[x - 1] + s1, '-' + s2)
elif u:
trace(delta, zeta, t, r, str1, str2,
x, y - 1, '-' + s1, str2[y - 1] + s2)
else:
r.append((s1, s2))
```

```
def B(zeta, delta):
n = len(X)
t = [[0 for _ in range(n+1)] for _ in range(n+1)]
for i in range(n+1):
t[0][i] = delta/2 * i
for i in range(n+1):
t[i][0] = delta/2 * i
for y in range(1, n+1):
for x in range(1, n+1):
v = X[x - 1] == Y[y - 1]
t[y][x] = max(
t[y][x - 1] + delta/2,
t[y - 1][x] + delta/2,
t[y - 1][x - 1] + 1 if v else t[y - 1][x - 1] + zeta)
score = t[n+1 - 1][n+1 - 1]
trace(delta, zeta, t, results, X, Y, x, y)
results = []

#This is optional. It prints out all optimal alignment
#for i, result in enumerate(results):
#print("Result {0}:\n{1}\n{2}"
#.format(i + 1, result[0], result[1]))

return score/n

#For printing out B at delta and zeta
print(B(zeta,delta))
```

Graphs

The next algorithm is used for sketching various graphs. Since B_n is piecewise affine, we can use bisection method's analogy to find values in which B_n is not differentiable. If we can calculate B_n at those values, we can sketch the graph of the function. Listing II sketches the following functions' graphs: B_n (as a variable of δ or ζ) and q_n.

Listing II: Graph

```
import matplotlib.pyplot as plt
#Fixed error parameter
epsilon = 10**(-5)

#Sketching functions B and p in relation to delta

def graph_B_delta(BM,BN,M,N,X,Y, epsilon):
T = (N+M)/2
BT=B(zeta,T)
if abs((BN+BM)/2-BT) > epsilon:
graph_B_delta(BM,BT,M,T,X,Y, epsilon)
graph_B_delta(BT,BN,T,N,X,Y, epsilon)
else:
plt.plot([M,T,N],[BM,BT,BN],'k')

def graph_q_delta(BM,BN,M,N,X,Y, epsilon):
T = (N+M)/2
q = (BM-BN)/(M-N)
BT=B(zeta,T)
if abs((BN+BM)/2-BT) > epsilon:
graph_q_delta(BM,BT,M,T,X,Y, epsilon)
graph_q_delta(BT,BN,T,N,X,Y, epsilon)
else:
plt.plot([M, N],[q, q],'k')

#Displaying the graph with pyplot
M = -7
N = 2
#BM = B(zeta,M)
#BN = B(zeta,N)
#plt.text(0.1,1.03,r'$y=q_{n}(\zeta^*,\delta)$'
  #,fontsize=14)
#plt.ylabel('$y$')
#plt.xlabel('$\delta$')
#plt.plot([M,N],[0,0],'k--')
#plt.plot([0,0],[BM-1,BN],'k--')
```

```
#plt.ylim([-0.2,1.2])
#plt.xlim([M,N])

#Sketching functions B and p in relation to zeta

def graph_B_zeta(BE,BF,E,F,X,Y,epsilon):
T = (E+F)/2
BT=B(T,delta)
if abs((BE+BF)/2-BT) > epsilon:
graph_B_zeta(BE,BT,E,T,X,Y,epsilon)
graph_B_zeta(BT,BF,T,F,X,Y,epsilon)
else:
plt.plot([E,T,F],[BE,BT,BF],'k')

def graph_p_zeta(BM,BN,M,N,X,Y,epsilon):
T = (N+M)/2
q = (BM-BN)/(M-N)
BT=B(T,delta)
if abs((BN+BM)/2-BT) > epsilon:
graph_p_zeta(BM,BT,M,T,X,Y,epsilon)
graph_p_zeta(BT,BN,T,N,X,Y,epsilon)
else:
plt.plot([M, N],[q, q],'k')

#Displaying the graph with pyplot
E = -1.5
F = 12
#BE = B(E, delta)
#BF = B(F, delta)
#plt.text(1, 0.83, r'$y=p_{n}(\zeta,\delta^*)$',
  #fontsize=14)
#plt.ylabel('$y$')
#plt.xlabel('$\zeta$')
#plt.plot([E,F],[0,0],'k--')
#plt.plot([0,0],[BE-1,BF+1],'k--')
#plt.ylim([-0.2,1.2])
#plt.xlim([E,F])

#This is where we choose which graph we want to sketch
#graph_B_delta(BM,BN,M,N,X,Y,epsilon)
#graph_q_delta(BM,BN,M,N,X,Y,epsilon)
#graph_B_zeta(BE1,BE2,E1,E2,X,Y,epsilon)
#graph_q_zeta(BF1,BF2,F1,F2,X,Y,epsilon)
plt.show()
```

Simulations

In simulations we used the following listing to generate random sequences of n uniformly distributed elements from alphabet $\mathbb{A} = \{a, b\}$.

Listing III: Random sequences

```
import random
import string
import sys

sys.setrecursionlimit(20000)

def getCode(length = 10, char = string.ascii_uppercase
+string.digits + string.ascii_lowercase ):
return ''.join(random.choice(char) for x in range(length))

n=1000

X = getCode(n, "ab")
Y = getCode(n, "ab")
```

Bibliography

[Bil] P. Billingsley, *Probability and Measure*, Third edition, A Wiley-Interscience Publication, New York, 1995.

[BSS] D. Fernández-Baca, Timo Seppäläinen, Giora Slutzki, *Bounds for Parametric Sequence Comparison*, 6th International Symposium on String Processing and Information Retrieval, Page 5, 1999.

[CZ] K. Chao, L. Zhang, *Sequence Comparison*, Springer, London, 2009.

[DGG] Luc Devroye, Gabor Lugosi, Laszlo Gyorfi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[FS] W. M. Fitch, T. F. Smith, *Optimal Sequence Alignments*, Proc. Nat. Acad. Sci., vol 80, pp. 1382-1386, 1983.

[GBN] D. Gusfield, K. Balasubramanian, D. Naor, *Parametric Optimization of Sequence Alignment*, Algorithmica, vol 12, pp. 312-326, 1994.

[LMT] J. Lember, H. Matzinger, F. Torres, *Proportion of Gaps and Fluctuations of the Optimal Score in Random Sequence Comparison*, Eichelsbacher P., Elsner G., Kösters H., Löwe M., Merkl F., Rolles S. (eds) Limit Theorems in Probability, Statistics and Number Theory. Springer Proceedings in Mathematics and Statistics, vol 42, pp 207-234, Springer, Berlin, 2013.

[Roc] R. Tyrrell Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1997.

[SK] D. Sankoff, J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley, Boston, 1983.

[Lem] Jüri Lember, *Informatsiooniteooria*, Lecture notes, 2011, https://www.math.ut.ee/sites/default/files/ms/informatsiooniteooria_2011.pdf

Non-exclusive licence to reproduce thesis and make thesis public

I, Rasmus Erlemann,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

    1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

    1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright, Effects of Parameters Choice in Random Sequence Comparison, supervised by Jüri Lember,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 03.08.2017