LARGE DEVIATION BASED UPPER BOUNDS FOR THE LCS-PROBLEM

Raphael Hauser [*], Servet Martinez [†] and Heinrich Matzinger [‡]

June 21, 2006

Abstract

Let $X := (X_1, \ldots, X_n)$ and $Y := (Y_1, \ldots, Y_n)$ be two finite sequences. Let L_n designate the length of the longest sequence which occurs as a subsequence of X as well as of Y. We analyze and apply a large deviation and Montecarlo simulation based method for the computation of improved upper bounds on the Chvàtal-Sankoff constant γ, which is defined by the limit $\gamma = \lim_{n\to\infty} \mathbb{E}[L_n]/n$ when X and Y are random sequences with i.i.d. entries. Our theoretical results show that this method converges to the exact value of γ when a control parameter m converges to infinity. We also give upper bounds on the complexity for numerically computing γ to any given precision via this method. Our numerical experiments confirm the theory and allow us to give new upper bounds that are correct to two digits.

AMS Classification: primary 05A16, 62F10; secondary 92E10.
Key Words: Longest common subsequence problem, Chvátal-Sankoff constant, upper bound, large deviation theory, Montecarlo simulation.

1 Introduction

The investigation of longest common subsequences (LCS) of two finite words is one of the main problems in the theory of pattern matching and plays a role in DNA- and Protein-

[*]Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, United Kingdom; Email: hauser@comlab.ox.ac.uk. This author's research is supported through a grant of the Nuffield Foundation under the "Newly Appointed Lecturers" grant scheme, project number NAL/00720/G.

[†]CMM-DIM-CNRS 2071, Universidad de Chile, Casilla 170-3 Correo 3 Santiago, Chile. E-mail: smartine@dim.uchile.cl. The research of this author is supported by Nucleus Millennium and Randomness P01-005.

[‡]Fakultät für Mathematik; Universität Bielefeld; D-33501 Bielefeld; Germany; Email: matzing@mathematik.uni-bielefeld.de. Also, School of Mathematics; Georgia Institute of Technology; 686 Cherry Street; Atlanta, GA 30332-0160; USA; Email: matzi@math.gatech.edu. This author thanks partial support from Nucleus Millennium P01-005 for his visit to CMM-DIM at Santiago.

alignments, file-comparison, speech-recognition and so forth.

Let $X := (X_1, \ldots, X_n)$ and $Y := (Y_1, \ldots, Y_n)$ be two independent randomly generated sequences with uniform i.i.d. entries from a finite alphabet $A = \{1, 2, \ldots, C\}$. In the simplest case the entries of X and Y are just i.i.d. Bernoulli variables with parameter $1/2$. Let L_n designate the length of a longest common subsequence of X and Y, that is, a sequence which occurs as a subsequence of both X and Y and which is of maximal length among all sequences with this property. The thus defined random variable L_n and several of its variants have been studied intensively by probabilists, computer-scientists and mathematical biologists; for applications of LCS-algorithms in biology see Waterman [26]. The books of Sankoff-Kruskal [23, 20], Capocelli [12, 13] and Apostolico-Crochemore-Galil-Manbar [3] present further applications.

Using a subadditivity argument, Chvàtal-Sankoff [14] prove that the limit

$$\gamma := \lim_{n\to\infty} \mathbb{E}[L_n]/n$$

exists. The exact value of γ remains however unknown. Chvàtal-Sankoff [14] derive upper and lower bounds for γ, and similar upper bounds were found by Baeza-Yates, Gavalda, Navarro and Scheihing [10] using an entropy argument. These bounds have been improved by Deken [17], and subsequently by Dančik-Paterson [15, 22]. In this paper we present a Monte Carlo and large deviation based method which allows to further improve the upper bounds on γ. Our approach can be seen as a generalization of the method of Dančik-Paterson.

The most widely used method for the comparison of genetic data is a generalization of the LCS-method. For an excellent overview of this subject see Waterman-Vingron [28]. In this generalization a maximal score is sought over the set of all possible alignments of the two sequences, where gaps are penalized with a fixed parameter $\delta > 0$ and mismatches are penalized by a fixed amount $\mu > 0$: consider for example the two words "brot" and "bat". One possible alignment $\mathbb{A}$ of these words is

b	r	o	t
b	a	$-$	t

The score of this alignment is $1 - \mu - \delta + 1 = S(\mathbb{A})$. The matching pairs of letters "b" and "t" are each valued with a weight of 1. The gap $-$ in "bat" after the "a" costs $-\delta$. Furthermore, the mismatch between "r" and "a" is penalized by adding $-\mu$ to the total score. If $M_{\mu,\delta}(X, Y)$ denotes the maximal score amongst all possible alignments of two words X and Y, and if $M_n(\mu, \delta)$ is the random variable defined by $M_n(\mu, \delta) = M_{\mu,\delta}(X, Y)$, where X and Y are two i.i.d. random sequences of length n, then the LCS-problem is a

special case of the investigation of $M_n(\mu, \delta)$, because $L_n = M_n(\infty, 0)$. Generalizing the arguments from the LCS-problem, one can prove that the limit

$$a(\mu, \delta) = \lim_{n \to \infty} \frac{\mathbb{E}[M_n]}{n}$$

exists. Arratia-Waterman [8] showed that there is a phase transition phenomenon defined by critical values of μ and δ. In one phase M_n is of linear order in n, whereas in the other it is logarithmically small in n. Waterman [27] conjectures that the deviation of M_n from its mean behaves like $\sqrt{n}$.

As mentioned earlier, the approach we use in this paper to derive upper bounds on γ is inspired by the method of Dančik-Paterson [15, 22]. However, in contrast to the latter, our method can be used in principle to derive upper bounds on $a(\mu, \delta)$ for values of μ and δ that correspond to the linear phase. This is a subject we plan to pursue in future research.

Let us mention a few further details on the history of these problems and the state of knowledge about them: Waterman-Arratia [8] derive a law of large deviation for L_n for fluctuations on scales larger than $\sqrt{n}$. The order of magnitude of the deviation from the mean of L_n is unknown, and in fact it is not even known if these deviations are larger than a power of n. However, using first passage percolation methods, Alexander [2] proves that $\mathbb{E}[L_n]/n$ converges at a rate of order $\sqrt{\log n/n}$.

Waterman [27] studies the statistical significance of the results produced by sequence alignment methods. An important problem that was open for decades concerns the longest increasing subsequence (LIS) of random permutations and appears to be related to the LCS-problem. However, it is an open question to know if solutions of the LIS-problem can be used to study the LCS problem, see Johansson [11] and Aldous-Diaconis [1].

Another problem related to the LCS-problem is that of comparing sequences X and Y by looking for longest common words that appear both in X and Y, and generalizations of this problem where the word does not need to appear in exactly the same form in the two sequences. The distributions that appear in this context have been studied by Arratia-Gordon-Goldstein-Waterman [4] and Neuhauser [21]. A crucial role is played by the Chen-Stein Method for the Poisson-Approximation. Arratia-Gordon-Waterman [5, 6] shed some light on the relation between the Erdös-Rényi law for random coin tossing and the above mentioned problem. In [7] the same authors also developed an extreme value theory for this problem.

2 Overview

As mentioned above, Dančik-Paterson [15, 22] derived the best deterministic bounds on the Chvàtal-Sankoff constant γ, that is, the numbers they derive are analytically proven

to be lower and upper bounds on γ respectively.

The results presented here are fundamentally different: we will derive a randomized algorithm that produces an upper bound $\hat{q}$ on γ at a given confidence level. For example, on the 95% level this means that $\mathbb{P}[\hat{q} > \gamma] \geq 0.95$. Thus, $\hat{q}$ is a random variable and a bound that is *not deterministic* but *probabilistic*. Moreover, $\hat{q}$ depends on the number l_0 of simulations and on a control parameter m whose role is further described below. For now it suffices to know that in each of the l_0 simulations we need to evaluate the length of the LCS of two random sequences of length $O(m)$ via the Wagner-Fischer algorithm [24] and collect certain information that is obtained "for free" from intermediate results during the computation. In our theoretical analysis we then show that $\hat{q}$ is a consistent estimator of γ, that is, $\lim_{m,l_0\to\infty} \hat{q} = \gamma$ almost surely. In fact, we show that asymptotically $\mathbb{P}[\gamma < \hat{q} < \gamma + \Xi] \geq \Lambda$ where $\Xi = O(m^{-\frac{\alpha}{2}})$ and $\Lambda = 1 - O(l_0^{-1})$, where $\alpha \in (0, 1)$ is a constant.

Ours are not the first results on simulated bounds that are consistent estimators of γ: Alexander [2] described a method that turns Montecarlo estimates $\bar{L}_n/n$ of $\mathbb{E}[L_n]/n$ into consistent upper and lower bounds of γ. Again, these bounds depend on the number l_0 of simulations and on the control parameter n, and it is the case that $\lim_{n,l_0\to\infty} \hat{q} = \gamma$ almost surely. Moreover, the midpoint $\hat{\gamma}_n$ between the upper and lower bounds determined by this method satisfies $\mathbb{P}[|\hat{\gamma}_n - \gamma| < \Xi] \geq \Lambda$ where $\Xi = O(n^{-1/2}) + O(l_0^{-1/2})$ and $\Lambda = 1 - \exp\big(-(O(n) + O(l_0))\big)$.

From a big-picture viewpoint the two methods thus appear to have similar properties. Note however that the above-mentioned convergence rates are asymptotic worst-case bounds obtained by analytic means and do not necessarily accurately describe the practical convergence behavior. There are therefore at least two strong motivations for analyzing the new approach:

(i) The new method is conceptually very different from Alexander's approach. This opens up a new class of algorithms with possible extensions to other related problems, in particular those appearing in connection with scoring functions in bioinformatics.

(ii) Practical versions of our algorithm converge orders of magnitude faster than the theoretical analysis predicts: with $m = 1000$ our method finds substantially tighter upper bounds on γ than Alexander's approach yields with $n = 50000$. Since the dominant work per simulation is due to an application of the Wagner-Fischer algorithm, the per-simulation complexity of our algorithm is $O(m^2)$ whereas that of Alexander's method is $O(n^2)$. Thus, from a practical point of view our method constitutes a considerable improvement.

3 Some Useful Notation and a Key Inequality

Let A be a finite alphabet and $A^* = \bigcup_{n\in\mathbb{N}} A^n$ be the set of finite words. We denote by $|A|$ the cardinal number of A, that is the number of symbols of the alphabet. For $a \in A^*$ denote by $|a|$ its length, that is, the number of letters in a. Trivially, $|ab| = |a| + |b|$ for every pair $(a, b) \in A^* \times A^*$, where ab denotes the concatenation of a and b, that is, the string consisting of the letters of a followed by those of b.

Let Π^n be the class of increasing sequences of $\{1, \dots, n\}$. We denote the cardinality of any $\pi \in \Pi^n$ by $|\pi|$, and its consecutive components by $\pi(i)$ $(i = 0, \dots, |\pi|$. For $a \in A^*$ and $\pi \in \Pi^{|a|}$ we use the notation $a_\pi := (a_{\pi(i)} : i = 1, \dots, |\pi|)$. The main object of study in this paper is the quantity

$$L(a, b) = \max\{k : \exists \pi \in \Pi^{|a|}, \sigma \in \Pi^{|b|}, |\pi| = k = |\sigma|, a_\pi = b_\sigma\},$$

that is, $L(a, b)$ is the length of a longest common subsequence of a and b.

For the analyis it is convenient to use the set of elementary events $\Omega = A^{\mathbb{N}} \times A^{\mathbb{N}}$ endowed with the canonical product σ-algebra. We will also sometimes identify Ω with $(A \times A)^{\mathbb{N}}$, and we denote the points of Ω by $\omega = (x, y)$, where $x = (x_n : n \in \mathbb{N})$ and $y = (y_n : n \in \mathbb{N})$. We use the following notation for the canonical projections defined on Ω: $X(\omega) = x$, $X_i(\omega) = x_i$, $Y(\omega) = y$ and $Y_j(\omega) = y_j$.

We endow Ω with a probability measure $\mathbb{P} = \mathbf{P} \times \mathbf{P}$, where $\mathbf{P}$ is a *Bernoulli measure* on $A^{\mathbb{N}}$, that is, $\mathbf{P} = \xi^{\mathbb{N}}$ where ξ is a probability distribution on the finite alphabet A with $\xi(a) > 0$ for all $a \in A$. In other words yet, all entries in X and Y are i.i.d. random variables with values in A and distribution ξ.

Remark 3.1. *It is interesting to note that some of the results presented in this paper extend to the situation where $\mathbb{P}$ is a ergodic shift-invariant measure on Ω. For example, the proof of relation* (3.1) *below goes through unchanged, and the argument we will present in* (3.3) *extends to the more general probability model if Birkhoff's Ergodic Theorem is invoked. However, since most of the results we present in this paper rely on $\mathbb{P}$ being a Bernoulli measure, and since this is the model of interest in the vast majority of applications, we decided keep to this slightly more restrictive framework.*

Let $x[i, j] = (x_k : i \le j)$ be the word formed by the letters between i-th and j-th coordinate on x. We use the same notation for words in y and for random vectors, that is, we write for example $X[i, j] = (X_k : i \le k \le j)$. Any pair of words $(a, b) \in A^* \times A^*$ defines a measurable set as follows,

$$[[a, b]] = \{(x, y) \in \Omega : x[1, |a|] = a, y[1, |b|] = b\}.$$

Extending this notation, we write $[[S]] = \cup_{(a,b)\in S}[[a,b]]$ for all $S \subseteq A^* \times A^*$.

Let $\left\{L_j^i : \Omega \to \mathbb{N} | i,j \in \mathbb{N}\right\}$ be the family of random variables

$$L_j^i = \begin{cases} L(X[i,j], Y[i,j]) & \text{if } i \le j, \\ 0 & \text{otherwise.} \end{cases}$$

For ease of notation we will write L_j for L_j^1. Then $\{L_j^i\}$ satisfies the hypotheses of Kingman's subadditive ergodic theorem, which implies that

$$\inf_{n\ge 1} \frac{L_n}{n} = \lim_{n\to\infty} \frac{L_n}{n} = \lim_{n\to\infty} \frac{\mathbb{E}[L_n]}{n} := \gamma \tag{3.1}$$

holds $\mathbb{P}$ almost everywhere on Ω for some real number γ, see e.g. [19]. The limit γ, trivially seen to be lying in the interval $(0,1)$, is called the *Chvàtal-Sankoff constant* associated with the law $\mathbb{P}$. It follows from (3.1) that for any $q < \gamma$ it is true that $\lim_{n\to\infty} \mathbb{P}\{L_n \ge qn\} = 1$. Therefore, for all $q \in (0,1)$,

$$\lim_{n\to\infty} \mathbb{P}\{L_n \ge qn\} < 1 \Rightarrow q \ge \gamma. \tag{3.2}$$

We write $S_1^n(q) := \{(a,b) \in A^n \times A^n : L(a,b) \ge qn\}$. Note that

$$\{L_n \ge qn\} = [[S_1^n(q)]] = \cup_{(a,b)\in S_1^n(q)}[[a,b]].$$

This notation will be useful in the proof of Lemma 3.1.

It will sometimes be necessary to have a lower bound for γ. An elementary relation is obtained as follows,

$$\gamma \overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{L_n}{n} \ge \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_k = Y_k} \overset{\text{a.s.}}{=} \sum_{a\in A} \mathbf{P}([[a]])^2 \ge |A|^{-1}. \tag{3.3}$$

The following definition introduces one of the key concepts upon which our methods rely:

Definition 3.1. *For any word $a \in A^*$ of length $|a| \ge 1$ let $a^- := (a_1, .., a_{|a|-1})$ be the word obtained by removing the last letter from a. For $m \in \mathbb{N}$ we say that a pair $(a,b) \in A^* \times A^*$ is a m-match if*

$$\begin{aligned} L(a,b) &= m, \\ L(a^-,b) &= m-1 \\ L(a,b^-) &= m-1. \end{aligned}$$

We write $\mathcal{M}^m$ for the set of m-matches in $A^* \times A^*$.

It follows immediately from Definition 3.1 that

$$(a,b) \in \mathcal{M}^m \Rightarrow \min\{|a|,|b|\} \geq m, \tag{3.4}$$

$$(X[1,i],Y[1,j]) \in \mathcal{M}^m, k \neq j \Rightarrow (X[1,i],Y[1,k]) \notin \mathcal{M}^m. \tag{3.5}$$

The last relation holds point-wise on Ω and says that, for a given $i \in \mathbb{N}$ there is at most one index j such that $(X[1,i],Y[1,j])$ is a m-match.

The following family of random variables will play an important role throughout all parts of this paper:

$$\begin{aligned} L_{i,j} &= L(X[1,i],Y[1,j]), \\ Z_{i,j}^{[m]} &= Z_{i,j} = \mathbf{1}_{\mathcal{M}^m}(X[1,i],Y[1,j]), \\ Z_k^{[m]} &= Z_k = \sum_{(i,j):i+j=k} Z_{i,j}. \end{aligned}$$

We will often use the simplified notation $Z_{i,j}, Z_k$ in contexts where we treat m as a fixed parameter. It follows immediately from (3.4) and (3.5) that $0 \leq Z_k \leq (k-2m)_+$, that is, $Z_k = 0$ everywhere on Ω for $k < 2m$. Associated with the variables Z_k is the following measure on $\mathbb{N}$ which will play a key role throughout our analysis: we set

$$\nu^{[m]}(k) = \nu(k) = \mathbb{E}[Z_k] \tag{3.6}$$

for all $k \in \mathbb{N}$, and ν is then extended to $\mathbb{N}$ by σ-additivity. Note that the definitions of $Z_{i,j}, Z_k$ and ν all depend on the choice of the parameter m. In order to avoid index cluttering we chose not to account for this dependence explicitly in the notation. This should not lead to confusion, but the reader should bear the dependence on m in mind. Let us mention that, although we cannot exclude at this point that ν be an infinite measure, we will later prove that it is finite because $\nu(\mathbb{N}) \leq |A|m$, see Lemma 4.3. However, ν is of course generally not a probability measure. A trivial identity which is sometimes useful is the following,

$$\nu(k) = \sum_{i+j=k} \mathbb{P}\{L_{i,j} = m\}. \tag{3.7}$$

We are ready to prove one of a key inequality that drives our approach:

Lemma 3.1. *Let $m \in \mathbb{N}$, $q \in [0,1]$, and let $\nu^{*\lfloor qn/m \rfloor}$ be the measure ν, defined in* (3.6)*, convoluted $\lfloor qn/m \rfloor$ times with itself. Then*

$$\mathbb{P}\{L_{n-1} \geq qn\} \leq \sum_{l_1+\cdots+l_{\lfloor qn/m \rfloor} \leq 2n} \nu(l_1) \cdot \cdots \cdot \nu(l_{\lfloor qn/m \rfloor}) = \nu^{*\lfloor qn/m \rfloor}\big([0,2n]\big). \tag{3.8}$$

Proof. Let us consider the class of words

$$S_2^n(q) = \cup_{(i,j):i+j=2n}\{(a,b) \in A^i \times A^j : L(a,b) \geq qn\}.$$

It is clearly the case that $S_1^n(q) \subseteq S_2^n(q)$. Let

$$S_3^{n,m}(q) := \Big\{ \big(a^1 \dots a^{\lfloor qn/m \rfloor} c^1, b^1 \dots b^{\lfloor qn/m \rfloor} c^2\big) : (a^k, b^k) \in \mathcal{M}^m (k = 1, \dots, \lfloor qn/m \rfloor), \\ c^1, c^2 \in A^*, \ \sum_{k=1}^{\lfloor qn/m \rfloor} |a^k b^k| + |c^1 c^2| = 2n \Big\}.$$

We claim that $S_2^n(q) \subseteq S_3^{n,m}(q)$. In fact, for any pair $(a,b) \in S_2^n(q)$, there exist two strictly increasing maps $\pi : [1, \lceil qn \rceil] \to [1, |a|]$ and $\sigma : [1, \lceil qn \rceil] \to [1, |b|]$ such that $a_\pi = b_\sigma$, and it is possible to choose π and σ minimal in the sense that for each pair $(\hat{\pi}, \hat{\sigma}) \in \Pi^{|a|} \times \Pi^{|b|}$ that satisfies

$$\begin{aligned} |\pi| = |\sigma| &= \lceil qn \rceil, \\ a_{\hat{\pi}} &= b_{\hat{\sigma}}, \\ \hat{\pi}(k) &\leq \pi(k) \quad (k = 1, \dots, \lceil qn \rceil), \\ \hat{\sigma}(k) &\leq \sigma(k) \quad (k = 1, \dots, \lceil qn \rceil), \end{aligned}$$

we have $\hat{\pi} = \pi$ and $\hat{\sigma} = \sigma$. It is easy to see that when π and σ are minimal in this sense, then

$$(a^k, b^k) := \big(a_{\pi(m(k-1)+1)} \dots a_{\pi(mk)}, b_{\sigma(m(k-1)+1)} \dots b_{\sigma(mk)}\big) \in \mathcal{M}^m$$

for $k = 1, \dots, \lfloor qn/m \rfloor$. Therefore, $\big(a^1 \dots a^{\lfloor qn/m \rfloor} c^1, b^1 \dots b^{\lfloor qn/m \rfloor} c^2\big) \in S_3^n(q)$, where $c^1 := a_{\pi(\lfloor qn/m \rfloor + 1)} \dots a_{|a|}$ and $c^2 := b_{\sigma(\lfloor qn/m \rfloor + 1)} \dots b_{\sigma(|b|)}$. This shows that $S_2^n(q) \subseteq S_3^{n,m}(q)$, as claimed.

It is now useful to introduce the index set

$$\mathcal{I}(q,n,m) = \Big\{ \vec{l} := (l_1, \dots, l_{\lfloor qn/m \rfloor}) \in \mathbb{N}^{\lfloor qn/m \rfloor} : \ \sum_{k=1}^{\lfloor qn/m \rfloor} l_k \leq 2n \Big\}. \tag{3.9}$$

With any element $\vec{l} \in \mathcal{I}(q,n,m)$ we associate the set

$$S_3^{n,m}(q, \vec{l}) := \Big\{ \big(a^1 \dots a^{\lfloor qn/m \rfloor} c^1, b^1 \dots b^{\lfloor qn/m \rfloor} c^2\big) \in S_3^{n,m}(q) : \\ |a^k b^k| = l_k, (k = 1, \dots, \lfloor qn/m \rfloor) \Big\}.$$

It is then clearly the case that

$$S_3^{n,m}(q) = \bigcup_{\vec{l}\in\mathcal{I}(q,n,m)} S_3^{n,m}(q,\vec{l}),$$

and hence that

$$\mathbb{P}([[S_3^{n,m}(q)]]) \leq \sum_{\vec{l}\in\mathcal{I}(q,n,m)} \mathbb{P}([[S_3^{n,m}(q,\vec{l})]])$$

which in turn implies

$$\begin{aligned}\mathbb{P}([[S_3^{n,m}(q)]]) &\leq \sum_{\vec{l}\in\mathcal{I}(q,n,m)} \sum_{S_3^n(q,\vec{l})} \mathbb{P}([[(a^1\dots a^{\lfloor qn/m\rfloor}c^1, b^1\dots b^{\lfloor qn/m\rfloor}, c^2)]]) \\ &\leq \sum_{\vec{l}\in\mathcal{I}(q,n,m)} \sum_{(a^k,b^k)\in\mathcal{M}^m:|a^kb^k|=l_k,(k=1,\dots,\lfloor qn/m\rfloor)} \prod_{k=1}^{\lfloor qn/m\rfloor} \mathbb{P}([[a^k,b^k]]),\end{aligned}$$

where the last inequality follows from the assumption that $\mathbb{P}$ is a Bernoulli measure and from the trivial inequality $\mathbb{P}([[c^1,c^2]]) \leq 1$. Now, since

$$\nu^{*\lfloor qn/m\rfloor}([0,2n]) = \sum_{\vec{l}\in\mathcal{I}(q,n,m)} \prod_{k=1}^{\lfloor qn/m\rfloor} \nu(l_k), \tag{3.10}$$

and

$$\prod_{k=1}^{\lfloor qn/m\rfloor} \nu(l_k) = \sum_{(a^k,b^k)\in\mathcal{M}^m:|a^kb^k|=l_k,(k=1,\dots,\lfloor qn/m\rfloor)} \prod_{k=1}^{\lfloor qn/m\rfloor} \mathbb{P}([[a^k,b^k]]),$$

we can conclude that

$$\mathbb{P}([[S_3^{n,m}(q)]]) \leq \nu^{*\lfloor qn/m\rfloor}([0,2n]). \tag{3.11}$$

Finally, since $\{L_n \geq qn\} = [[S_1^n(q)]] \subseteq [[S_2^n(q)]] \subseteq [[S_3^{n,m}(q)]]$, the proof is complete. □

4 A Large Deviation Based Upper Bound on γ

In this section we will apply large deviation techniques to find the exponential rate of the bound on the right hand side of (3.11). Since ν is not a probability measure in general, we will derive the relevant inequalities without using the classical results stated for probability distributions. Using the usual measure theoretic notation, we have

$$\Big(\int_{\mathbb{N}} \mathrm{e}^{t\left(\frac{2m}{q}-x\right)}\, d\nu(x)\Big)^{\lfloor qn/m\rfloor} = \sum_{(l_1,\dots,l_{\lfloor qn/m\rfloor})\in\mathbb{N}^{\lfloor qn/m\rfloor}} \mathrm{e}^{t\sum_{k=1}^{\lfloor qn/m\rfloor}\left(\frac{2m}{q}-l_k\right)} \prod_{k=1}^{\lfloor qn/m\rfloor} \nu(l_k).$$

Since every $(l_1, \ldots, l_{\lfloor qn/m \rfloor}) \in \mathcal{I}(q,n,m)$ satisfies $\sum_{k=1}^{\lfloor qn/m \rfloor}(2m/q - l_k) \geq -2m/q$, (3.10) implies

$$\nu^{*\lfloor qn/m \rfloor}\big([0,2n]\big) \leq \Big(\int_{\mathbb{N}} \mathrm{e}^{t\left(\frac{2m}{q}-x\right)}\, d\nu(x)\Big)^{\lfloor qn/m \rfloor} \mathrm{e}^{\frac{2mt}{q}}\,. \tag{4.1}$$

This leads to the following theorem, providing the main tool for the construction of our upper bounds on γ:

Theorem 4.1. *Let $t > 0$ and $q \in [0,1]$. If*

$$\sum_{k\in\mathbb{N}} \mathrm{e}^{t\left(\frac{2m}{q}-k\right)}\, \nu(k) < 1 \tag{4.2}$$

then $\gamma < q$.

Proof. If (4.2) holds then for all n large enough the right hand side of (4.1) is < 1. The result then follows from (3.2) and (3.8). □

Let us now define

$$q_1(m) := \inf\Big\{q \in [0,1] :\ \exists t > 0 \text{ s.t. } \sum_{k\in\mathbb{N}} \mathrm{e}^{t\left(\frac{2m}{q}-k\right)}\, \nu(k) < 1\Big\}. \tag{4.3}$$

By Theorem (4.1) we it is then true that $\gamma \leq q_1(m)$ for all $m \in \mathbb{N}$. In the remainder of this section, culminating in Theorem 4.3 below, we will show that $\lim_{m\to\infty} q_1(m) = \gamma$. The analysis that leads to this result also sets the stage for understanding the practical Montecarlo methods to compute $q_1(m)$ devised in Section 5. We start by recalling the following large-deviation inequality:

Lemma 4.1 (Azuma-Hoeffding). *Let $t \in \mathbb{N}$, $\mathcal{F} = \cup_{s\in\mathbb{N}_0}\mathcal{F}_s$ a filtration and $V_0, V_1, \ldots, V_t$ a $\mathcal{F}$-adapted martingale such that $V_0 = 0$. Let $a > 0$ and $\Delta > 0$, and let us assume that for all $s \in [0, t-1]$ it is the case that $|V_t - V_{t+1}| \leq a$ a.s. Then the following inequality holds true,*

$$\mathbb{P}\big\{V_t \geq \Delta t\big\} \leq \mathrm{e}^{-\frac{t\Delta^2}{2a^2}}$$

Proof. This result is due to Azuma [9] and Hoeffding [18]. A modern proof can be found for example in [25], Section 11.1.4. □

We will now use Lemma 4.1 to show that $L_{i,j}$ decays exponentially:

Lemma 4.2. *For all $\Delta \geq 0$ it is true that $\mathbb{P}\big\{L_{i,j} \geq \frac{i+j}{2}(\gamma+\Delta)\big\} \leq \mathrm{e}^{-(i+j)\Delta^2/8}$.*

Proof. We have $L_{i+j,j+i} \geq L_{i,j} + L_{j,i} \circ (\sigma_X^i, \sigma_Y^j)$, where σ_X and σ_Y denote the left-shift operators on the X and Y components of (X, Y) respectively. Since $\mathbb{P}$ is a Bernoulli measure, $L_{i,j}$ and $L_{j,i} \circ (\sigma_X^i, \sigma_Y^j)$ are identically distributed, so that $\mathbb{E}[L_{i+j,j+i}] \geq 2\mathbb{E}[L_{i,j}]$. It follows from subadditivity that $\mathbb{E}[L_{i+j,j+i}] \leq \gamma(i+j)$, implying $\mathbb{E}[L_{i,j}] \leq \gamma(i+j)/2$ and hence,

$$\mathbb{P}\Big\{L_{i,j} \geq \frac{i+j}{2}(\gamma+\Delta)\Big\} \leq \mathbb{P}\Big\{L_{i,j} \geq \mathbb{E}[L_{i,j}] + \frac{(i+j)}{2}\Delta\Big\}. \tag{4.4}$$

Let us next consider a fixed path $\Gamma : \{0, \ldots, i+j\} \to \mathbb{Z}^2$ that leads from $\Gamma(0) = (0,0)$ to $\Gamma(i+j) = (i,j)$ by moving one unit in the positive direction of either coordinate in each step. Let $r(k)$ and $s(k)$ be defined by $G(k) = \big(r(k), s(k)\big)$, let $\mathcal{F}_0 = \{\mathbb{R}, \emptyset\}$ be the trivial σ-algebra on $\mathbb{R}$, and let

$$\mathcal{F}_k = \sigma\left(X_u, Y_v : u = 1, \ldots, r(k); v = 1, .., s(k)\right), \quad (k = 1, \ldots, i+j).$$

(Here and elsewhere the notation extends in a natural way to the case where an index set is empty. For example, if $r(k) = 0$ then $\mathcal{F}_k = \sigma\big(Y_1, \ldots, Y_{s(k)}\big)$.) For $k \in \{0, \ldots, i+j\}$ let us define $V_k := \mathbb{E}\big[L_{i,j} - \mathbb{E}[L_{i,j}]\big\|\mathcal{F}_k\big]$.

The sequence $V_0, V_1, \ldots, V_{i+j}$ is then a martingale that satisfies the conditions of Lemma 4.1 with $a = 1$. Applying the lemma, we obtain the inequality

$$\mathbb{P}\left(L_{i,j} - \mathbb{E}[L_{i,j}] \geq \frac{(i+j)}{2}\Delta\right) \leq \mathrm{e}^{-(i+j)\Delta^2/8}.$$

Combined with (4.4) this yields the result. □

Remark 4.1. *Applying the Azuma-Hoeffding Lemma to the martingale $(-V_0, \ldots, -V_{i+j})$, where V_k is as in the proof of Lemma 4.2, one finds the inequality*

$$\mathbb{P}\left(L_{i,j} - \mathbb{E}[L_{i,j}] \leq -\frac{(i+j)}{2}\Delta\right) \leq \mathrm{e}^{-(i+j)\Delta^2/8}. \tag{4.5}$$

As a consequence of Lemma 4.2 we can now bound $\nu(k)$ for small k:

Corollary 4.1. *Let $k \leq 2m/\gamma$ and $\Delta'_k = (2m/k) - \gamma$. Then*

$$\nu(k) \leq 2m|A|\,\mathrm{e}^{-(\Delta'_k)^2 k/8}\,.$$

Proof. Let us consider a pair (i,j) such that $k := i + j \leq 2m/\gamma$. Then $\Delta'_k \geq 0$, and we can apply Lemma 4.2 to find that

$$\mathbb{P}(L_{i,j} = m) \leq \mathbb{P}(L_{i,j} \geq m) \leq \mathrm{e}^{-(\Delta'_k)^2 k/8}\,.$$

Together with (3.7) and (3.3) this proves the claim. □

As promised in Section 3, we will next prove that ν is a finite measure. Recall again that the definitions of $Z_{i,j}$, Z_k and ν depend on the value of the control parameter m.

Lemma 4.3. *For every $m \in \mathbb{N}$, it is true that*

$$\sum_{k\geq 1} \nu(k) = \mathbb{E}\left[\sum_{i,j>0} Z_{i,j}\right] \leq |A|m.$$

Proof. The sequence $(\mathcal{Z}_m : m \in \mathbb{N})$ of random variables

$$\mathcal{Z}_m := \min\{k \geq 0 : Z_{m,k}^{[m]} = 1\}$$

is strictly increasing in m. Moreover, we have $\mathcal{Z}_1 = \min\{k \geq 1 : Y_k = X_1\}$. Hence,

$$\mathbb{P}\{\mathcal{Z}_1 = k\} = \sum_{a\in A} \xi(a)(1-\xi(a))^{k-1}\xi(a), \tag{4.6}$$

and we find that $\mathbb{E}[\mathcal{Z}_1] = |A|$.

Next, let us set $\mathcal{Y}_0 = 0$ and $\mathcal{Y}_k = \min\{l > \mathcal{Y}_{k-1} : Y_l = X_k\}$ for $k \geq 1$. Then $\mathcal{Y}_1 = \mathcal{Z}_1$ and $\mathcal{Y}_k \geq \mathcal{Z}_k$ holds true for all $k \in \mathbb{N}$. Because $\mathbb{P}$ is a Bernoulli measure, $\mathcal{Y}_{k+1} - \mathcal{Y}_k$ is independent of $\big(\mathcal{Y}_l : l < k\big)$ and is identically distributed as $\mathcal{Y}_1$. Therefore, we have

$$\mathbb{E}[\mathcal{Z}_m] \leq \mathbb{E}[\mathcal{Y}_m] \leq m\mathbb{E}[\mathcal{Z}_1] = m|A|. \tag{4.7}$$

Let us now consider the random index set

$$\mathbf{M}^m = \{(i,j) \in \mathbb{N} : (X[1,i], Y[1,j]) \in \mathcal{M}^m\}$$

corresponding to the m-matches occurring in X and Y. Since $(m, \mathcal{Z}_m) \in \mathbf{M}^m$, it follows from (3.4), (3.5) and the definition of an m-match that

$$|\mathbf{M}^m| = \sum_{i,j>0} Z_{i,j} \leq \mathcal{Z}_m - m < \sum_{k=1}^{m} \mathcal{W}_k, \tag{4.8}$$

where $\mathcal{W}_k = \mathcal{Y}_k - \mathcal{Y}_{k-1}$ $(k = 1, \ldots, m)$ are i.i.d. random variables distributed according to (4.6). Therefore, by virtue of (4.7) we obtain

$$\mathbb{E}\big[\sum_{i,j>0} Z_{i,j}\big] \leq \mathbb{E}[\mathcal{Z}_m - m] \leq |A|(m-1),$$

proving the claim. □

Corollary 4.2. *For all $a \in A$ let $\eta(a) = (1-\xi(a))^{1/m}$. Then for all $k \in \mathbb{N}$ the following holds true,*

$$\nu(k) \leq \big(\max_{a\in A} \eta(a)\big)^{k-2} \sum_{a\in A} m\xi(a) \frac{k-(k-1)\eta(a)}{\big(1-\eta(a)\big)^2}.$$

Proof. We use the notation and facts derived in the proof of Lemma 4.3. Note that $Z_k^{[m]} > 0$ implies that $k \leq \mathcal{Z}_m$. Therefore,

$$\begin{aligned}
\nu(k) = \nu^{[m]}(k) &= \mathbb{E}\big[Z_k^{[m]}\big] = \sum_{r=1}^{\infty} \mathbb{P}\big(Z_k^{[m]} \geq r\big) \\
&= \sum_{s=m}^{\infty} \sum_{r=1}^{s} \mathbb{P}\big(Z_k^{[m]} \geq r \| \mathcal{Z}_m = s\big) \cdot \mathbb{P}\big(\mathcal{Z}_m = s\big) \\
&\leq \sum_{s=k}^{\infty} s\mathbb{P}\big(\sum_{l=1}^{\infty} \mathcal{W}_l \geq s\big) \leq \sum_{s=k}^{\infty} s \sum_{l=1}^{m} \mathbb{P}\Big(\mathcal{W}_l > \frac{s-1}{m}\Big) \\
&= \sum_{s=k}^{\infty} sm \sum_{a\in A} \xi(a)\big(1-\xi(a)\big)^{\lfloor \frac{s-1}{m} \rfloor} \leq \sum_{a\in A} \frac{m\xi(a)}{\eta(a)} \sum_{s=k}^{\infty} s\eta(a)^{s-1} \\
&= \sum_{a\in A} \eta(a)^{k-2} \cdot m\xi(a) \cdot \frac{k-(k-1)\eta(a)}{\big(1-\eta(a)\big)^2}.
\end{aligned}$$

□

Our next result is instrumental in proving the consistency of the estimator $q_1(m)$:

Theorem 4.2. *Let $\Delta > 0$ be such that $q = \Delta + \gamma \leq 1$, and let $0 < t \leq \frac{\Delta}{8|A|^2}$. Then*

$$\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)}\, \nu(k) \leq \left(m|A| + 4m^2|A|^2\right) \mathrm{e}^{-t\Delta m} \tag{4.9}$$

Proof. It follows from the hypotheses that $1/\gamma = \Delta/(\gamma q) + 1/q$. Thus, $1/\gamma \geq \Delta + 1/q$ and

$$a := \frac{2m}{q} + m\Delta < \frac{2m}{\gamma} < 2m|A|, \tag{4.10}$$

where the last inequality follows from (3.3). We split the left hand side of (4.9) as follows,

$$\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)}\, \nu(k) = \sum_{k<a} \mathrm{e}^{t(2m/q-k)}\, \nu(k) + \sum_{k\geq a} \mathrm{e}^{t(2m/q-k)}\, \nu(k), \tag{4.11}$$

and we derive bounds on both right-hand terms separately.

To bound the second term, note that for $k \geq a$ we have $2m/q - k \leq 2m/q - a = -\Delta m$. Therefore,

$$\sum_{k\geq a} \mathrm{e}^{t(2m/q-k)}\, \nu(k) \leq \mathrm{e}^{-t\Delta m} \sum_{k\geq a} \nu(k) \leq m|A|\, \mathrm{e}^{-t\Delta m}, \tag{4.12}$$

where the second inequality follows from the fact that Lemma 4.3 implies that $\sum_{k\geq a} \nu(k) \leq \sum_{k\geq 1} \nu(k) \leq m|A|$.

To bound the first term in (4.11), note that (3.4) implies $\nu(k) = 0$ for $k < 2m$. Using this in conjunction with (4.10) and Corollary 4.1 we find

$$\begin{aligned} \sum_{k<a} \mathrm{e}^{t(2m/q-k)}\, \nu(k) &\leq 2m|A| \sum_{k=2m}^{a-1} \mathrm{e}^{t(2m/q-k)}\, \mathrm{e}^{-(\Delta'_k)^2 k/8} \\ &\leq 2m|A| \sum_{k=2m}^{a-1} \mathrm{e}^{t(2m/q-k)}\, \mathrm{e}^{-(\Delta'_k)^2 m/4}, \end{aligned} \tag{4.13}$$

where $\Delta'_k := 2m/k - \gamma$, and where the last inequality holds because $k \geq 2m$. If we now use the change of variables $\bar{k} := a - k$, then

$$\sum_{k=2m}^{a-1} \mathrm{e}^{t(2m/q-k)}\, \mathrm{e}^{-(\Delta'_k)^2 m/4} = \mathrm{e}^{-tm\Delta} \sum_{\bar{k}=1}^{a-2m} \mathrm{e}^{t\bar{k}}\, \mathrm{e}^{-(\Delta''_{\bar{k}})^2 m/4}, \tag{4.14}$$

where

$$\Delta''_{\bar{k}} := \frac{2m}{a-\bar{k}} - \gamma = \frac{2m}{a}\frac{1}{(1-\bar{k}/a)} - \gamma \geq \frac{2m}{a} - \gamma + \frac{2m\bar{k}}{a^2}.$$

Note that

$$\frac{2m}{a} - \gamma = \frac{1}{\frac{1}{q}+\frac{\Delta}{2}} - \gamma = \frac{q}{1+\frac{\Delta q}{2}} - \gamma \geq q\Big(1-\frac{\Delta q}{2}\Big) - \gamma \geq \Delta - \frac{\Delta q^2}{2} \geq \frac{\Delta}{2}.$$

Together with (4.10) this yields

$$(\Delta'_{\bar{k}})^2 \geq \left(\frac{\Delta}{2} + \frac{\bar{k}}{2m|A|^2}\right)^2 > \frac{\Delta\bar{k}}{2m|A|^2}. \tag{4.15}$$

Substituting (4.15) into (4.14), we get

$$\begin{aligned}\sum_{k=2m}^{a-1} \mathrm{e}^{t(2m/q-k)}\,\mathrm{e}^{-(\Delta'_k)^2 m/4} &\overset{(4.15)}{\leq} \mathrm{e}^{-tm\Delta}\sum_{\hat{k}=1}^{a-2m}\mathrm{e}^{t\bar{k}-\frac{\Delta\bar{k}}{8|A|^2}} \leq \mathrm{e}^{-tm\Delta}(a-2m)\\ &\overset{(4.10)}{<} 2m|A|\,\mathrm{e}^{-tm\Delta},\end{aligned} \tag{4.16}$$

where the second inequality is a consequence of the hypothesis on t. The result now follows from (4.11), (4.12), (4.13) and (4.16). □

We are finally ready to prove that $q_1(m)$ is consistent in m:

Theorem 4.3. *If $q_1(m)$ is defined as in (4.3), then*

$$\lim_{m\to\infty} q_1(m) = \gamma.$$

Proof. Because of Theorem 4.1 we already know that $q_1(m) \geq \gamma$ for all $m \in \mathbb{N}$. The result will thus be shown if we can prove that

$$\limsup_{m\to\infty} q_1(m) \leq \gamma. \tag{4.17}$$

Let $\epsilon > 0$ be fixed, and let us choose Δ and t as a function of m as follows: $\Delta := m^{-1/(2+\epsilon)}$ and $t =: \Delta/(8|A|^2)$. Then for all m large enough, the conditions of Theorem 4.2 are satisfied. Moreover, we have

$$\mathrm{e}^{-t\Delta m} = \mathrm{e}^{-\frac{\Delta^2 m}{8|A|^2}} = \mathrm{e}^{-\frac{m^{\frac{\epsilon}{2+\epsilon}}}{8|A|^2}},$$

so that, again for m large enough, $\big(2m + 4m^2|A|^2\big)\,\mathrm{e}^{-t\Delta m} < 1$. Theorem 4.2 thus implies that there exists $m_0 \in \mathbb{N}$ such that for all $m \geq m_0$,

$$\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)}\,\nu(k) < 1, \tag{4.18}$$

and then $q_1(m) \leq \gamma + \Delta = \gamma + m^{-1/(2+\epsilon)}$ by (4.3), showing that (4.17) is indeed true. □

5 Montecarlo Simulation

Theorem 4.1 revealed that whenever $0 \leq q \leq 1$ and $t > 0$ are such that

$$\sum_{k\in\mathbb{N}} \mathrm{e}^{t(2m/q-k)}\,\nu(k) < 1, \tag{5.1}$$

then q is an upper bound on the Chvàtal-Sankoff constant γ. When using this theoretical tool in practical calculations one faces the problem that one cannot evaluate (5.1) explicitly, because $\nu(k)$ is not known except for very small values of k. A practical way to get around this problem is to evaluate (5.1) via Montecarlo simulation.

Recall that we assumed that $\{X_i\}_{\mathbb{N}} \cup \{Y_j\}_{\mathbb{N}}$ is a family of i.i.d. random variables which take values in the finite alphabet A according to a probability law ξ. Recall also that in Section 3 we introduced the notation

$$Z_{i,j} = \mathbf{1}_{\mathcal{M}^m}\big(X[1,i], Y[1,j]\big), \quad Z_k = \sum_{i+j=k} Z_{i,j}, \quad \text{and} \quad \nu(k) = \mathbb{E}[Z_k].$$

Let us now define the random variable

$$W = W(t,q) := \sum_{k>0} \mathrm{e}^{t(2m/q-k)}\,Z_k, \tag{5.2}$$

so that $\mathbb{E}[W] = \sum_{k>0} \mathrm{e}^{t(2m/q-k)}\,\nu(k)$ is the expression of interest in (5.1).

For the purposes of Montecarlo simulation, we consider $(X_i^l : i,l \in \mathbb{N})$ and $(Y_j^l : j,l \in \mathbb{N})$, two independent collections of i.i.d. random variables with distribution ξ on A. Let us define

$$\begin{aligned} Z_{i,j}^l &:= \mathbf{1}_{\mathcal{M}^m}\big(X^l[1,i], Y^l[1,j]\big), \\ Z_k^l &:= \sum_{i+j=k} Z_{i,j}^l, \quad \text{and} \\ W^l = W^l(t,q) &:= \sum_{k>0} \mathrm{e}^{t(2m/q-k)}\,Z_k^l. \end{aligned}$$

Thus, Z_k^l counts the number of m-matches of length k observed in the l-th realization (X^l, Y^l) of the pair of random sequences. Then

$$\hat{\nu}_k := \frac{1}{l_0} \sum_{l=1}^{l_0} Z_k^l$$

is an unbiased estimator of $\nu(k)$ and

$$\frac{1}{l_0} \sum_{l=1}^{l_0} W^l = \sum_{k>0} \mathrm{e}^{t(2m/q-k)}\, \hat{\nu}_k \tag{5.3}$$

is an unbiased estimator of the left hand side of (5.1).

In Section 5.1 we will show how the estimator (5.3) can be used in theory to obtain an upper bound on γ to any given precision and at any given confidence level. In Section 5.2 we will also derive an upper bound on the number of elementary computer operations necessary to compute such a bound as a function of the required precision and confidence level. The theoretical analysis is based on estimates which are unnecessarily conservative in practice. Practical implementations are therefore based on a slightly different approach, leading to a number of issues that need careful attention. We discuss these in Section 5.3.

5.1 Montecarlo Simulation in Theory

The main result of this section is the following theorem, which gives a tool to determine the value of the control parameter m and the number l_0 of simulations necessary to obtain an estimator $\hat{q}_1$ of γ to within a specific precision and on a given confidence level:

Theorem 5.1. *Let $\alpha, \delta \in (0,1)$ be constants and $l_0 \in \mathbb{N}$. Let us choose Δ and t as functions of m as follows: $\Delta = m^{-\alpha/2}$ and $t_1 = \Delta/(16|A|^2)$. Let us finally consider*

$$\hat{q}_1 = \Delta + \inf\Big\{ q > 0 : \sum_{k=1}^{\infty} \mathrm{e}^{t_1(2m/q-k)}\, \hat{\nu}_k < 1 \Big\} \tag{5.4}$$

as an estimator for q_1. Then there exists a number

$$m_0 = m_0(\alpha, \xi, \delta) < \max\left(O\left(\Big(\frac{2}{1-\gamma} \Big)^{\frac{2}{\alpha}} \right), O\left(\Big(|A|^4 \log \frac{2}{\delta} \Big)^{\frac{1+\epsilon}{1-\alpha}} \big(1 - \log \min_{a \in A} \xi(a) \big) \right) \right),$$

where ϵ is a small number, such that for all $m \geq m_0$ it is true that

$$\mathbb{P}\big(\gamma \leq \hat{q}_1 \leq \gamma + 2\Delta \big) \geq 1 - \mathrm{e}^{-l_0(1-\delta)^2/2} - \frac{8}{l_0}. \tag{5.5}$$

Note that the right hand side of (5.5) determines the confidence level that the computed estimator $\hat{q}_1$ *is* an upper bound *and* approximates γ to within precision 2Δ. The confidence level increases if the number l_0 of simulations is increased. The precision on the other hand increases with m. α and δ merely play the role of control parameters. These could be fixed at given values, but treating them as parameters reveals how their values affect the complexity estimates of Section 5.2. The same pertains to the dependence of m_0 on the distribution ξ on A. Note finally that γ and $|A|$ are functions of ξ, which is why no extra variables are necessary in $m_0(\alpha, \xi, \delta)$.

Before we can prove Theorem 5.1 we need three preliminary results. The first lemma shows that when m is large enough, then with high probability there will be a m-match of length not much larger than what γ predicts:

Lemma 5.1. *Let α and Δ be as in Theorem 5.1 and let us consider the event*

$$B := \left\{ \exists i, j \ \textit{s.t.}\ Z_{i,j} = 1 \ \textit{and}\ i + j \leq 2 \left\lceil \frac{m}{\gamma} + \frac{\Delta m}{2} \right\rceil \right\}. \tag{5.6}$$

Then there exists a number $m_1 = m_1(\alpha, \xi)$ such that for all $m \geq m_1$,

$$\mathbb{P}(B) \geq 1 - \exp\Big(-\frac{m^{1-\alpha}|A|^{-4}}{256}\Big). \tag{5.7}$$

Proof. Alexander [2] showed that there exists a constant $C > 0$,independent of ξ and A, such that for all $n \geq 1$,

$$0 \leq \gamma - \frac{\mathbb{E}[L_n]}{n} \leq C\sqrt{\frac{\log n}{n}}.$$

A more explicitly quantitative version of this result can be obtained as follows: by choosing $\lambda = 2$, $\theta = 3$ in Proposition 2.4 of [2], a relaxation of Equation (2.13) in [2] shows that

$$0 < \gamma - \frac{\mathbb{E}[L_n]}{n} < 7\sqrt{\frac{\log n}{n}} \qquad \forall n \geq 16. \tag{5.8}$$

Let $k' = m/\gamma + \Delta m/2$ and $n' = \lceil k' \rceil$. Then $m \geq 16$ implies

$$n' \geq k' > m \geq 16. \tag{5.9}$$

Moreover, if

$$m \geq m_2(\alpha, \xi) := \inf\Big\{ y > 0 : \ \frac{x^{1-\alpha}}{2 \cdot 56^2 |A|^3} > \log x, \ \ \forall x \geq y \Big\}$$

then (3.3) implies

$$\log m < \frac{m^{1-\alpha}\gamma^3}{2 \cdot 56^2}. \tag{5.10}$$

Finally, if

$$m \geq m_3(\alpha, |A|) := \big(2 \cdot 56^2 |A|^3 \log(|A| + 1/2)\big)^{\frac{1}{1-\alpha}}$$

then (3.3) implies

$$\frac{m^{1-\alpha}\gamma^3}{2 \cdot 56^2} > \log\Big(\frac{1}{\gamma} + \frac{1}{2}\Big). \tag{5.11}$$

Now, for

$$m \geq m_4(\alpha, |A|) := \max\big(m_2(\alpha, |A|), m_3(\alpha, |A|)\big),$$

(5.10) and (5.11) show that

$$\begin{aligned}
\log k' = \log\Big(\frac{m}{\gamma} + \frac{\Delta m}{2}\Big) &< \log m + \log\Big(\frac{1}{\gamma} + \frac{1}{2}\Big) \\
&\overset{(5.10),(5.11)}{<} \frac{m^{1-\alpha}\gamma^3}{56^2} \\
&< \frac{m^{-\alpha}\gamma^4}{56^2}\Big(\frac{m}{\gamma} + \frac{m^{-\frac{\alpha}{2}}m}{2}\Big) \\
&= \Big(\frac{\Delta\gamma^2}{8}\Big)^2 \frac{k'}{7^2},
\end{aligned} \tag{5.12}$$

and then (5.8), (5.9) and (5.12) show that for

$$m \geq m_5(\alpha, \xi) := \max\big(16, m_4\big)$$

the following holds,

$$0 < \gamma - \frac{\mathbb{E}[L_{n'}]}{n'} < 7\sqrt{\frac{\log n'}{n'}} \leq 7\sqrt{\frac{\log k'}{k'}} < \frac{\Delta\gamma^2}{8}. \tag{5.13}$$

Using the notation $\gamma_{n'} := \mathbb{E}[L_{n'}]/n'$, (5.13) and (3.3) imply

$$\gamma - \gamma_{n'} - \frac{\Delta\gamma^2}{4} \leq -\frac{\Delta\gamma^2}{8} \leq -\frac{\Delta|A|^{-2}}{8} \tag{5.14}$$

Now note that when the event $D := \big\{L_{n'} \geq m\big\}$ holds, then a m-match of total length less than or equal to n' must have occurred within $\big(X[1, n'], Y[1, n']\big)$. Hence, $D \subseteq B$, and it follows that

$$\mathbb{P}(\Omega \setminus B) \leq \mathbb{P}(\Omega \setminus D). \tag{5.15}$$

We have

$$\Omega \setminus D = \{L_{n'} < m\} = \Big\{ \frac{L_{n'}}{n'} - \gamma_{n'} < \frac{m}{n'} - \gamma_{n'} \Big\}. \tag{5.16}$$

Moreover, if

$$m \geq m_1(\alpha, \xi) := \max\big(m_5, 2^{\frac{2}{\alpha}}\big),$$

then $\Delta\gamma/2 < m^{-\frac{\alpha}{2}}/2 < 1/4$. Observe that for $x \in [0, 1/4]$ it is true that $1/(1+x) \leq 1 - x/2$. Applying this inequality to $x = \Delta\gamma/2$, we find

$$\frac{m}{n'} - \gamma_{n'} \leq \frac{m}{k'} - \gamma_{n'} = \frac{\gamma}{1 + \frac{\Delta\gamma}{2}} - \gamma_{n'} \leq \gamma - \gamma_{n'} - \frac{\Delta\gamma^2}{4}.$$

Substituting this into (5.16) yields

$$\mathbb{P}(\Omega \setminus D) \leq \mathbb{P}\Big(\frac{L_{n'}}{n'} - \gamma_{n'} \leq \gamma - \gamma_{n'} - \frac{\Delta\gamma^2}{4} \Big). \tag{5.17}$$

Combining the last inequality with (5.14), we find that

$$\mathbb{P}(\Omega \setminus D) \leq \mathbb{P}\Big(\frac{L_{n'}}{n'} - \frac{\mathbb{E}[L_{n'}]}{n'} \leq -\frac{\Delta |A|^{-2}}{8} \Big) \leq \mathrm{e}^{-n' \frac{\Delta^2 |A|^{-4}}{256}}, \tag{5.18}$$

where the last inequality follows from (4.5). Since $n' > m$, we find that the bound on the right hand side of (5.18) is smaller than $\exp(-m\Delta^2 |A|^{-4}/256)$. Finally, using $\Delta = m^{-\alpha/2}$, (5.15) and (5.18), we find that for all $m \geq m_1(\alpha, |A|)$,

$$\mathbb{P}(\Omega \setminus B) \leq \exp(-m^{1-\alpha} |A|^{-4}/256),$$

which is of course equivalent to the claimed inequality (5.7). □

Our second lemma shows that if m is large enough then the probability of finding an estimator value $\hat{q}_1$ significantly below γ is very small:

Lemma 5.2. *Let α, δ and Δ be as in Theorem 5.1, and let $\hat{\nu}_k := \sum_{l=1}^{l_0} Z_k^l / l_0$ for all $k \in \mathbb{N}$. Then there exists a number $m_6 = m_6(\alpha, \xi, \delta)$ such that for all $m \geq m_6$, $t \geq \Delta/(16|A|^2)$ and $q \in (0, \gamma - \Delta)$, it is true that*

$$\mathbb{P}\Big(\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q - k)}\, \hat{\nu}_k < 1 \Big) \leq \mathrm{e}^{-l_0 (1-\delta)^2/2}. \tag{5.19}$$

Proof. Note that $|\gamma - q| \geq \Delta$ implies $|2m/\gamma - 2m/q| \geq 2\Delta m$. Thus, when $q \leq \gamma - \Delta$, we find that $2m/q - k \geq \Delta m - 1$ for every $k \leq \lceil 2m/\gamma + \Delta m \rceil$. Hence,

$$\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)}\, Z_k \geq \sum_{k=1}^{\lceil 2m/\gamma+\Delta m\rceil} \mathrm{e}^{t(2m/q-k)}\, Z_k \geq \mathrm{e}^{t(\Delta m-1)} \sum_{k=1}^{\lceil 2m/\gamma+\Delta m\rceil} Z_k \geq \mathrm{e}^{t(\Delta m-1)}\, \mathbf{1}_B, \quad (5.20)$$

where $\mathbf{1}_B$ denotes the indicator function of the event B defined in (5.6). By definition,

$$\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)}\, \hat{\nu}_k = \sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)} \Big(\frac{1}{l_0} \sum_{l=1}^{l_0} Z_k^l\Big) = \frac{1}{l_0} \sum_{l=1}^{l_0} \sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)}\, Z_k^l.$$

It follows therefore from (5.20) that

$$\mathbb{P}\Big(\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)}\, \hat{\nu}_k < 1\Big) \leq \mathbb{P}\Big(\frac{1}{l_0} \sum_{l=1}^{l_0} \mathrm{e}^{t(\Delta m-1)}\, \mathbf{1}_B^l < 1\Big). \quad (5.21)$$

Here, $\big(\mathbf{1}_B^l : l \in \mathbb{N}\big)$ denotes an i.i.d. sequence of copies of the random variable $\mathbf{1}_B$.

Now, for all $m \geq m_7$, where

$$m_7 = m_7(\alpha, \xi, \delta) = \big(1 + 16|A|^2 \log(2/\delta)\big)^{\frac{1}{1-\alpha}},$$

we have

$$t(\Delta m - 1) \geq \frac{m^{1-\alpha} - m^{-\frac{\alpha}{2}}}{16|A|^2} > \frac{m^{1-\alpha} - 1}{16|A|^2} \geq \log(2/\delta),$$

and hence, $\mathrm{e}^{t(\Delta m-1)} > 2/\delta$. Moreover, it follows from Lemma 5.1 that if $m \geq m_8$, where

$$m_8 = m_8(\alpha, \xi, \delta) = \max\Big(m_1, \big(256|A|^4 \log(2/\delta)\big)^{\frac{1}{1-\alpha}}\Big),$$

then

$$\mathbb{E}\big[\mathbf{1}_B^l\big] = \mathbb{P}(B) \geq 1 - \exp\Big(-\frac{m^{1-\alpha}|A|^{-4}}{256}\Big) \geq 1 - \frac{\delta}{2}.$$

Therefore, for $m \geq m_6 := \max(m_7, m_8)$ we have

$$\Big(\frac{1}{l_0} \sum_{l=1}^{l_0} \mathrm{e}^{t(\Delta m-1)}\, \mathbf{1}_B^l < 1\Big) \Rightarrow \Big(\frac{1}{l_0} \sum_{l=1}^{l_0} \mathbf{1}_B^l < \frac{\delta}{2}\Big) \Rightarrow \Big(\frac{1}{l_0} \sum_{l=1}^{l_0} \big(\mathbf{1}_B^l - \mathbb{E}[\mathbf{1}_B^l]\big) \leq -(1-\delta)\Big). \quad (5.22)$$

Applying Lemma 4.1 with $a = 1$ to the martingale defined by

$$V_0 = 0, \qquad \mathcal{F}_0 = \{\emptyset, \mathbb{R}\},$$
$$V_k = \sum_{l=1}^{k} \big(\mathbb{E}[\mathbf{1}_B^l] - \mathbf{1}_B^l\big), \qquad \mathcal{F}_k = \sigma(V_1, \ldots, V_k), \quad (k = 1, \ldots, l_0),$$

we have

$$\mathbb{P}\Big(\frac{1}{l_0}\sum_{l=1}^{l_0}\big(\mathbb{E}[\mathbf{1}_B^l]-\mathbf{1}_B^l\big)\geq 1-\delta\Big)\leq \mathrm{e}^{-l_0(1-\delta)^2/2}\,.$$

Together with (5.21) and (5.22) this finishes the proof. □

The third lemma allows us to give a bound on $VAR(W)$ that will be needed in the proof of Theorem 5.1:

Lemma 5.3. *Let α and Δ be as in Theorem 5.1, and let $q=q(m)=\gamma+\Delta$. Then for all $m\geq m_9(\alpha,\xi):=(1-\gamma)^{-\frac{2}{\alpha}}$ and for all $t\in(0,\Delta/(16|A|^2)]$ it is true that*

$$\mathbb{E}\left[\Big(\sum_{k=1}^{\infty}\mathrm{e}^{t_1(2m/q-k)}\,Z_k\Big)^2\right]\leq\left(161m^4|A|^4+\frac{2m|A|}{\min_{a\in A}\xi(a)}\right)\mathrm{e}^{-2t\Delta m}\,. \tag{5.23}$$

Proof. Let $a=a(m):=2m/q+m$. (3.3) shows that $q>\gamma\geq|A|^{-1}$, and hence,

$$a<2m\big(|A|+1\big). \tag{5.24}$$

We will use the splitting

$$\mathbb{E}\left[\Big(\sum_{k=1}^{\infty}\mathrm{e}^{t_1(2m/q-k)}\,Z_k\Big)^2\right]\leq 2\mathbb{E}\left[\Big(\sum_{k\leq a}\mathrm{e}^{t_1(2m/q-k)}\,Z_k\Big)^2\right]+2\mathbb{E}\left[\Big(\sum_{k>a}\mathrm{e}^{t_1(2m/q-k)}\,Z_k\Big)^2\right] \tag{5.25}$$

and bound both terms on the right hand side separately.

When $k>a$, we have $2m/q-k<-m<-\Delta m$, and hence,

$$\sum_{k>a}\mathrm{e}^{t(2m/q-k)}\,Z_k\leq\mathrm{e}^{-t\Delta m}\sum_{k>0}Z_k\overset{(4.8)}{\leq}\mathrm{e}^{-t\Delta m}\sum_{k=1}^{m}\mathcal{W}_k,$$

where $\{\mathcal{W}_k:\ k=1,\dots,m\}$ are the i.i.d. random variables defined in the proof of Lemma 4.3 and whose distribution has the moment generating function

$$\Phi(s)=\sum_{a\in A}\frac{\xi(a)s}{1-s\big(1-\xi(a)\big)}.$$

This implies that

$$\begin{aligned}\mathbb{E}\left[\Big(\sum_{k>a} \mathrm{e}^{t(2m/q-k)}\, Z_k\Big)^2\right] &\le \mathrm{e}^{-2t\Delta m}\,\mathbb{E}\left[\Big(\sum_{k=1}^{m} \mathcal{W}_k\Big)^2\right] \\ &= \mathrm{e}^{-2t\Delta m}\left(m\mathbb{E}[\mathcal{W}_1^2] + 2\tbinom{m}{2}\mathbb{E}[\mathcal{W}_1]^2\right) \\ &= \mathrm{e}^{-t\Delta m}\big(m\Phi''(1) - m\Phi'(1) + m(m-1)\Phi'(1)^2\big) \\ &= \mathrm{e}^{-t\Delta m}\Big(2m\sum_{a\in A}\frac{1}{\xi(a)} - 3m|A| + m(m-1)|A|^2\Big). \end{aligned} \tag{5.26}$$

Note that $t\Delta m$ is a positive power of m with our choice of t and m.

On the other hand,

$$\begin{aligned}\Big(\sum_{k\le a} \mathrm{e}^{t(2m/q-k)}\, Z_k\Big)^2 &\le \Big(\sum_{k\le q} Z_k\Big)\Big(\sum_{k\le a} \mathrm{e}^{2t(2m/q-k)}\, Z_k\Big) \\ &\le a^2 \sum_{k>0} \mathrm{e}^{2t(2m/q-k)}\, Z_k, \end{aligned} \tag{5.27}$$

where the last inequality follows from $Z_k \le (k-2m)_+ \le a$. Since $2t \in (0, \Delta/8|A|]$ satisfies the conditions of Theorem 4.2, and since $q = \gamma + \Delta \le 1$ for all $m \ge m_9$, where

$$m_9 = m_9(\alpha, \xi) := (1-\gamma)^{-\frac{2}{\alpha}},$$

(4.9),(5.24) and (5.27) imply

$$\mathbb{E}\left[\Big(\sum_{k\le a} \mathrm{e}^{t(2m/q-k)}\, Z_k\Big)^2\right] \le 4m^3|A|\big(|A|+1\big)^2\big(1+4|A|m\big)\,\mathrm{e}^{-2t\Delta m}\,. \tag{5.28}$$

Using (5.25), (5.26) and (5.28), the result readily follows. □

We are finally ready to give a proof of Theorem 5.1:

Proof. Consider the events

$$E_{m,1} := \Big\{\sum_{k=1}^{\infty} \mathrm{e}^{t_1(2m/q-k)}\,\hat{\nu}_k \ge 1, \quad \forall q \in (0, \gamma-\Delta)\Big\}, \qquad \text{and}$$

$$E_{m,2} := \Big\{\sum_{k=1}^{\infty} \mathrm{e}^{t_1(2m/(\gamma+\Delta)-k)}\,\hat{\nu}_k < 1\Big\}.$$

Then (5.4) shows $E_{m,1} \subseteq \{\gamma \leq \hat{q}_1\}$ and $E_{m,2} \subseteq \{\hat{q}_1 \leq \gamma + 2\Delta\}$, which implies that

$$1 - \mathbb{P}\left(\gamma \leq \hat{q}_1 \leq \gamma + 2\Delta\right) \leq \mathbb{P}\left(\Omega \setminus E_{m,1}\right) + \mathbb{P}\left(\Omega \setminus E_{m,2}\right). \tag{5.29}$$

Lemma 5.2 provides a bound on the first term on the right hand side of this inequality, because it shows that for $m \geq m_6$,

$$\mathbb{P}\left(\Omega \setminus E_{m,1}\right) \leq \mathrm{e}^{-l_0(1-\delta)^2/2}. \tag{5.30}$$

To bound the second term on the right hand side of (5.29), let $W(t,q)$ be defined as in (5.2). Then $\mathbb{E}[W(t,\gamma+\Delta] = \sum_{k=1}^{\infty} \mathrm{e}^{t_1(2m/(\gamma+\Delta)-k)}\, \nu(k)$. By Chebychev's inequality,

$$\begin{aligned} &\mathbb{P}\left(\Big|\sum_{k=1}^{\infty} \mathrm{e}^{t_1(2m/(\gamma+\Delta)-k)}\, \hat{\nu}_k - \mathbb{E}[W(t_1,\gamma+\Delta)]\Big| \geq \frac{1}{2}\right) \\ &\quad = \mathbb{P}\left(\Big|\frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t_1,\gamma+\Delta) - \mathbb{E}[W(t_1,\gamma+\Delta)]\Big| \geq \frac{1}{2}\right) \leq \frac{4\mathbb{E}[W(t_1,\gamma+\Delta)^2]}{l_0}. \end{aligned} \tag{5.31}$$

Note that for all $m \geq m_9$, t_1 and Δ satisfy the conditions of Theorem 4.2 which shows that for all $m \geq \max(m_9, m_{10})$ with

$$m_{10} = m_{10}(\alpha,\xi) := \inf\Big\{y > 0 : \big(x|A| + 4x^2|A|^2\big)\, \mathrm{e}^{-\frac{x^{1-\alpha}}{16|A|^2}} \leq \frac{1}{2}, \quad \forall x \geq y\Big\}$$

it is true that $\mathbb{E}[W(t_1,\gamma+\Delta)] \leq 1/2$. Likewise, Lemma 5.3 shows that for all $m \geq \max(m_9, m_{11})$ with

$$m_{11} = m_{11}(\alpha,\xi) := \inf\left\{y > 0 : \left(161x^4|A|^4 + \frac{2x|A|}{\min_{a\in A}\xi(a)}\right)\mathrm{e}^{-\frac{x^{1-\alpha}}{16|A|^2}} \leq \frac{1}{2}, \quad \forall x \geq y\right\}$$

it is the case that $\mathbb{E}[W(t_1,\gamma+\Delta)^2] \leq 1/2$. But then (5.31) yields

$$\mathbb{P}\left(\Omega \setminus E_{m,2}\right) \leq \frac{8}{l_0}. \tag{5.32}$$

The inequalities (5.29), (5.30) and (5.32) show that the theorem holds true for $m_0(\alpha,\xi,\delta) = \max\big(m_6, m_9, m_{10}, m_{11}\big)$. The claim on the order of m_0 as a function of α, ξ and δ is easy to check directly. □

5.2 Theoretical Complexity of Montecarlo Simulation

In this paragraph we will briefly discuss the computational complexity for simulating an upper bound to a given precision and at a given confidence level. The analysis has already been done in Section 5.1, all that remains to do is to extract the information from the results we developed there.

Let $\Lambda \in (0,1)$ be a given confidence level and let $\Xi \in (0,1)$ be a given maximum permissible error. If we wish to simulate an estimate $\hat{q}_1$ that is an upper bound on and an approximation of the Chvàtal-Sankoff constant γ to within precision Ξ at the confidence level Λ, then how large do the control parameters m and l_0 have to be chosen to guarantee such an outcome? Theorem 5.1 shows that if

$$m \geq \max\left(\left(\frac{2}{\Xi}\right)^{\frac{2}{\alpha}}, m_0(\alpha,\xi,\delta)\right), \qquad l_0 \geq \frac{16}{1-\Lambda}, \qquad \delta = l_0^{-\frac{1}{4}}, \tag{5.33}$$

then

$$\mathbb{P}\left(\gamma \leq \hat{q}_1 \leq \gamma + \Xi\right) \geq \mathbb{P}\left(\gamma \leq \hat{q}_1 \leq \gamma + 2\Delta\right) \geq 1 - \mathrm{e}^{-l_0(1-\delta)^2/2} - \frac{8}{l_0} \geq 1 - \frac{16}{l_0} \geq \Lambda,$$

that is to say, (5.33) guarantees that $\hat{q}_1$ has the desired properties. Since α, ξ and δ are fixed, it is the first part of the term defining m that becomes dominant for small Ξ. The value of α that minimizes the required size of m depends on Ξ but is bounded away both from 0 and 1. Thus, if one works with the lower bounds on m and l_0 derived in (5.33), then

$$m = O\left(\Xi^{-\frac{2}{\alpha}}\right) \quad \text{and} \quad l_0 = O\left(\frac{1}{1-\Lambda}\right) \tag{5.34}$$

is required to compute an upper bound estimate $\hat{q}_1$ on the levels of accuracy Ξ and confidence Λ. This confirms what we have mentioned earlier: increasing m leads to better precision whereas increasing l_0 leads to a higher confidence level.

Let us now determine an estimate on the number of elementary computer operations required to perform the simulation of $\hat{q}_1$ with values of m and l_0 as given in (5.34): to construct $\hat{q}_1$, one needs to generate l_0 independent copies $\left(X^l, Y^l\right)$ of $(X,Y) = \big((X_1, X_2, \dots), (Y_1, Y_2, \dots)\big)$. In fact, each pair has to be generated only up to the finite random length that contains the full set of m-matches defined by $\left(X^l, Y^l\right)$. Lemma 4.3 and equation (4.8) show that we can expect that only about $m|A|$ terms have to be generated for each pair (X^l, Y^l) to account for all m-matches contained in it, and Corollary 4.2 implies that it is exponentially rare in m that more than $O(m)$ terms need to be generated. Computing the set of all m-matches contained in a pair (X^l, Y^l) takes therefore $O(m^2)$ computer operations when the Wagner-Fischer algorithm [24] is applied. Thus, generating all the m-matches that occur in the l_0 independent copies of (X,Y) takes $O(l_0 m^2)$

time. Since each pair $\big(X^l, Y^l\big)$ contains at most $3m$ m-matches, computing $\hat{\nu}$ and $\hat{q}_1$ from the generated data takes $O(l_0 m)$ time. The overall complexity for the simulation of $\hat{q}_1$ is therefore $O\big(l_0 m^2\big)$ operations. Because of (5.34), the complexity of computing an upper bound on γ at the precision level Ξ and confidence level Λ is therefore

$$O\left(\frac{1}{(1-\Lambda)\cdot \Xi^{\frac{4}{\alpha}}}\right) \tag{5.35}$$

operations.

Note that the complexity estimate (5.35) is an upper bound that corresponds to the worst case scenario. The practical complexity is lower, as we will see in Section 5.3. Note also that we did not specify what we mean by a "computer operation". In fact, our arguments are based on the assumption that a computer can perform operations with real numbers. We do not enter a discussion of round-off and finite-precision issues here, but taking these into account it is not difficult to see that (5.35) is also a complexity bound in terms of floating-point operations.

5.3 Montecarlo Simulation in Practice

Unfortunately, the complexity bound (5.35) is valid only asymptotically for very large values of m, because it is assumed that $m \geq m_0(\alpha, \xi, \delta)$: in fact, if $A = \{0, 1\}$ with ξ the standard Bernoulli measure characterized by $\xi(0) = 1/2 = \xi(1)$ (i.e., this is coin flipping), and if $\alpha = 0.1$ and $\delta = 0.1$ for example, then the complexity bound (5.35) only holds for $m \geq mm_0 = O(10^6)$, which is beyond reach in practical computations. Thus, while the complexity analysis of this section is interesting on theoretical grounds, practical methods cannot rely on it. In this section we will discuss practical methods that can achieve about two correct digits of accuracy with $m = 1000$ for the coin flipping example mentioned above. We will see that such practical methods pose a new set of challenges that need careful attention in the implementations.

Our theoretical analysis of Section 5.1 crucially depended on the fact that $\hat{q} \geq \gamma + \Delta$, where $\Delta \simeq O(m^{-\alpha/2})$. In essence, what we proved is that for m large enough and $\hat{q} \geq \gamma + \Delta$, the expression

$$\inf_{t>0} \sum_{k>0} \mathrm{e}^{t(2m/\hat{q}-k)}\, \hat{\nu}_k \tag{5.36}$$

is exponentially small in m and thus very close to zero, and that the probability that $\hat{q}_1$ lies in this range is exponentially small in the number l_0 of Montecarlo simulations.

Unfortunately, for $\alpha = 0.5$ and $m = 1000$ for example, this means that $\hat{q}_1 \geq \gamma + 0.1778$, which leads to an upper bound that is not satisfactorily close to the true value of γ. Therefore, in the practical use of the method, we would like to consider estimator values

$\hat{q}$ that are allowed to lie in the interval $(\gamma, \gamma + \Delta)$. In this case the expression (5.36) is smaller than 1 only by a small amount.

5.3.1 Setting the Stage for a Practical Algorithm

Since (5.36) is a random variable, the basic problem of the practical approach is to decide whether the sample of this variable obtained in a Montecarlo simulation lies significantly below 1 or not. An answer to this question is of course provided by Chebychev's inequality: we would like to design an estimator $\hat{q}$ such that the probability of wrongly deciding that $\hat{q}$ is an upper bound on γ is smaller than $1 - \Lambda$, that is, for the specific value of $t > 0$ used in the decision, we require that

$$\mathbb{P}\Big(\hat{q} < \gamma,\ \sum_{l=1}^{l_0} \mathrm{e}^{t(2m/\hat{q}-k)}\, \hat{\nu}_k < 1\Big) \le 1 - \Lambda \tag{5.37}$$

But Chebychev's inequality implies that

$$\begin{aligned}
&\mathbb{P}\Big(\hat{q} < \gamma,\ \sum_{l=1}^{l_0} \mathrm{e}^{t(2m/\hat{q}-k)}\, \hat{\nu}_k < 1\Big) \le \mathbb{P}\Big(\mathbb{E}\big[W(t,\hat{q})\big] > 1,\ \frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t,\hat{q}) < 1\Big) \\
&\le \mathbb{P}\Big(\Big|\frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t,\hat{q}) - \mathbb{E}\big[W(t,\hat{q})\big]\Big| > 1 - \frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t,\hat{q})\Big) \\
&\le \frac{VAR\big(W(t,\hat{q})\big)}{l_0\Big(1 - \frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t,\hat{q})\Big)^2}
\end{aligned}$$

Therefore, (5.37) is satisfied if

$$\frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t,\hat{q}) \le 1 - \sqrt{\frac{\hat{v}(t,\hat{q})}{(1-\Lambda)l_0}}, \tag{5.38}$$

where $\hat{v}(t,\hat{q})$ is an upper bound on $VAR\big(W(t,\hat{q})\big)$.

This leads to the problem of determining such an upper bound $\hat{v}(t,\hat{q})$. Note that (5.23) only applies for $q \ge \gamma + \Delta$ and hence is not useful in the practical context. A way out of this dilemma is to choose $\hat{v}(t,\hat{q})$ as a statistical estimator, defined in terms of the data $\{Z_k^l :\ k \ge 1, 1 \le l \le l_0\}$, such that

$$\mathbb{P}\left(VAR\big(W(t,\hat{q})\big) \le \widehat{v}(t,\hat{q})\right) \ge 1 - \eta \cdot (1 - \Lambda) \tag{5.39}$$

for some $\eta \in (0,1)$, and to accept $\hat{q}$ as an upper bound on γ if

$$\frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t,\hat{q}) \le 1 - \sqrt{\frac{\hat{v}(t,\hat{q})}{(1-\eta)(1-\Lambda)l_0}}, \tag{5.40}$$

is satisfied. This procedure yields an upper bound on γ at the confidence level at least Λ: the probability of wrongly deciding that $\hat{q}$ is an upper bound on γ is bounded as follows,

$$\begin{aligned}
&\mathbb{P}\Big(\hat{q} < \gamma,\ \sum_{l=1}^{l_0} \mathrm{e}^{t(2m/\hat{q}-k)}\,\hat{\nu}_k < 1\Big) \\
&\quad \le \mathbb{P}\Big(\hat{q} < \gamma,\ \sum_{l=1}^{l_0} \mathrm{e}^{t(2m/\hat{q}-k)}\,\hat{\nu}_k < 1, VAR\big(W^l(t,\hat{q})\big) \le \widehat{v}(t,\hat{q})\Big) + \mathbb{P}\Big(VAR\big(W^l(t,\hat{q})\big) > \widehat{v}(t,\hat{q})\Big) \\
&\quad \le (1-\eta)(1-\Lambda) + \eta(1-\Lambda) = 1-\Lambda.
\end{aligned}$$

Thus, the challenge is to design the estimators $\hat{q}$ and $\hat{v}(t,\hat{q})$ so that (5.39) and (5.40) are satisfied and $\hat{q}$ is as close to γ as possible, that is, as small as possible. Let us assume for a moment that the choice of $\hat{v}(t,q)$ has been fixed. Then a good choice for $\hat{q}$ is the solution of the following nonlinear optimization problem,

$$\begin{aligned}
\hat{q} &= \min_{(t,q)\in\mathbb{R}^2} q \\
\text{subject to } \frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t,q) &\le 1 - \sqrt{\frac{\hat{v}(t,q)}{(1-\eta)(1-\Lambda)l_0}} \\
q \ge 0, t &\ge 0.
\end{aligned} \tag{5.41}$$

Note that η, m, l_0 and the simulated data $\{Z_k^l :\ k \ge 1, 1 \le l \le l_0\}$ are all parameters that define (5.41), but when computing $\hat{q}$ we are interested in the situation where these parameters are fixed. Of course, the resulting value of $\hat{q}$ becomes a function of the parameters.

Let us now discuss how to define the estimator $\hat{v}(t,q)$. It follows from Corollary 4.2 that $\nu(k)$ is exponentially small in k for large k. Since moreover $\nu(k) = 0$ for $k < 2m$, this implies that $\mathbb{E}[W(t,q)]$, $VAR(W(t,q))$ and moments of all orders of $W(t,q)$ exist, suggesting the following approach: the empirical variance

$$\widehat{VAR}_{l_0}\big(W(t,q)\big) = \frac{1}{l_0-1}\sum_{k=1}^{l_0}\Big(W^k(t,q) - \frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t,q)\Big)^2, \tag{5.42}$$

is an unbiased estimator of $VAR\big(W(t,q)\big)$. Thus, if reasonable assumptions can be made about the distribution of the empirical variance (5.42) or a similar expression, then using a confidence interval argument one can define $\hat{v}$ so that it satisfies (5.39).

5.3.2 The Pitfalls of Variance Estimation

Before we put the outlined approach into practice, let us explain the pitfalls that need to be avoided. If $2m/\gamma - k > 0$ or, in other words, if k is small in comparison to the typical total length of a m-match, then Corollary 4.1 shows that

$$\nu(k) \leq 2m|A|\,\mathrm{e}^{-\frac{1}{8}\left(\frac{2m}{\gamma}-k\right)^2\frac{\gamma^2}{k}}\,.$$

This behavior is qualitatively correct: the first plot of Figure 1 shows the empirical distribution $\hat{\nu}$ obtained in $l_0 = 80000$ simulations for the coin flipping example and $m = 100$. The second plot shows that for k in the lower tail end $230 \leq k \leq 240$ the expression $-log\big(\nu(1/y)\big)$ as a function of $y = k^{-1}$ nearly coincides with the function $35000y - 142.6$. That is, for small k the measure ν nearly behaves like $\nu(k) \simeq \exp(a/k+b)$ with $a = -35000$ and $b = 142.6$. This leads to the following dilemma:

On the one hand, since $\mathbb{P}(Z_k^l > 0) \leq \mathbb{E}[Z_k^l] = \nu(k)$*, the event* $\{Z_k^l > 0\}$ *is exponentially rare in* $1/k$ *for* $2m \leq k < 2m/\gamma$*, and hence this occurs rarely in simulations.*

On the other hand, when $Z_k^l > 0$ *does occur then for* $q \simeq \gamma$ *the term* $\exp(t(2m/q - k))$ *is exponentially large in* $t > 0$ *for the same range of* k*, so that* $\exp(t(2m/q - k))\nu(k)$ *makes a nonnegligible contribution to* $W^l(t, q)$*.*

To make matters worse, when $Z_k^l > 0$ for some $k \in [2m, 2m/\gamma)$ then generally $Z_k^l > 0$ for other nearby values of k because the random variables $\{Z_k^l : k \in \mathbb{N}\}$ are not independent. Thus, unless t is very small, it could occur that most samples of $W^l(t, q)$ lie around a cluster of small values, while a few outliers take significantly larger values. Such a situation renders accurate estimates of $VAR(W)$ by (5.42) extremely costly in terms of the number of simulations required. Figure 2 illustrates that this phenomenon occurs not only in theory, but also in practice: the plots show histograms of 50000 samples of $W^l(t, q)$ for the coin flipping example with $m = 100$, $q = 0.825$, and for $t = 0.1, 0.2, 0.3$ and 0.4 respectively. Note that for $t \geq 0.2$ small clusters of large values are observed.

Whether or not the observed outlier points affect the estimation of $VAR(W)$ depends on their probability weights. An investigation of the distribution tails of $W(t, q)$ is therefore revealing. Of course, we already noted that the tails of ν decrease exponentially, and that $\nu(k) = 0$ for $k < 2m$. These two properties assure that for fixed t and q the distribution tails of $W(t, q)$ decay exponentially, that is, $\mathbb{P}\big(W(t, q) > \tau\big) \leq b\,\mathrm{e}^{-a\tau}$ for some constants $a, b > 0$ and $\tau \gg 1$. Note however that this exponential decay may become effective only for very large τ, so that from an empirical point of view the decay might be algebraic, that is,

$$\mathbb{P}\big(W(t, q) > \tau\big) \sim \mathrm{e}^b \cdot \tau^{-a} \tag{5.43}$$

for some constants $a > 0$, $b \in \mathbb{R}$. Figure 3 shows that this is indeed the case in the above

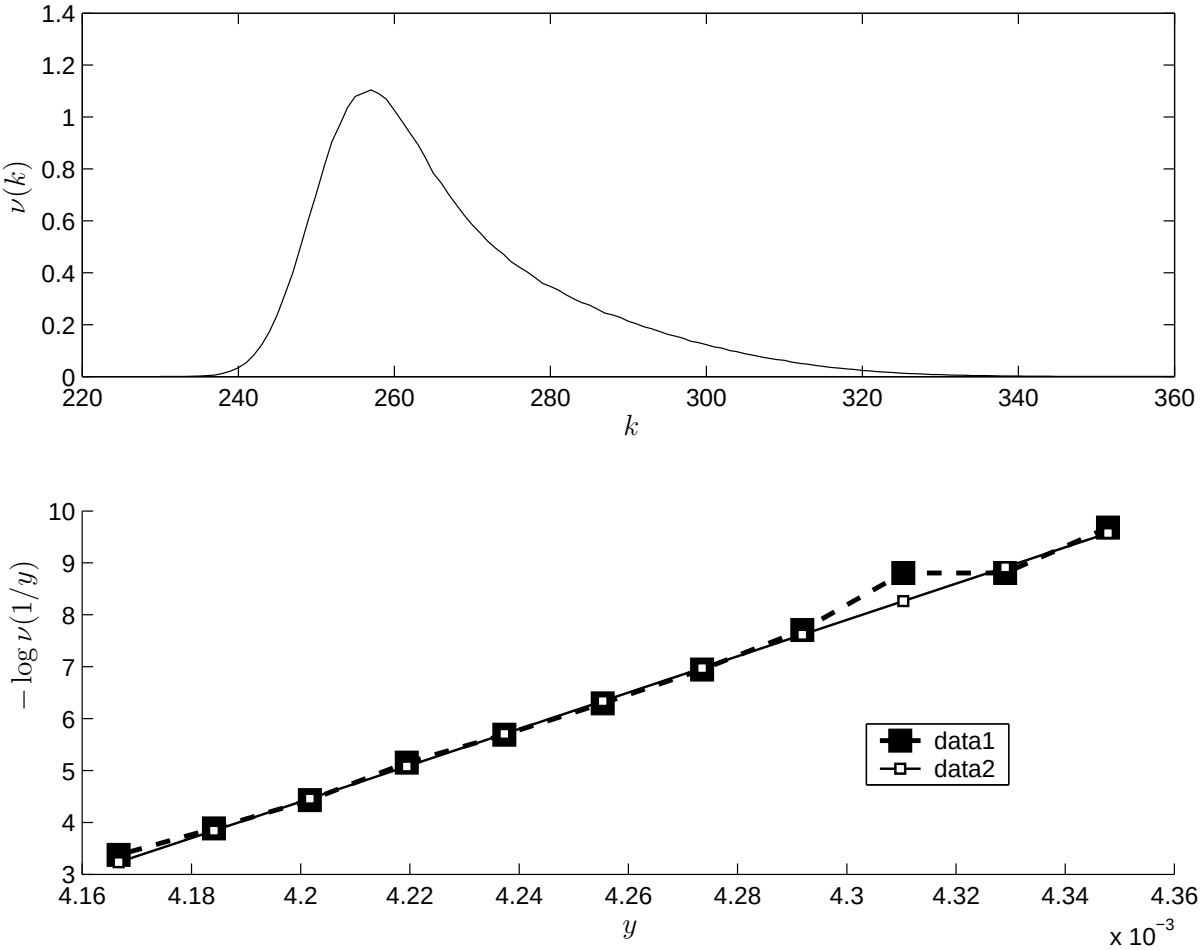

Figure 1: $\hat{\nu}$ and its lower tail end for $m = 100$ and $l_0 = 80000$. Data1 represent the function $-log\big(\nu(1/y)\big)$ and data2 the function $35000y - 142.6$.

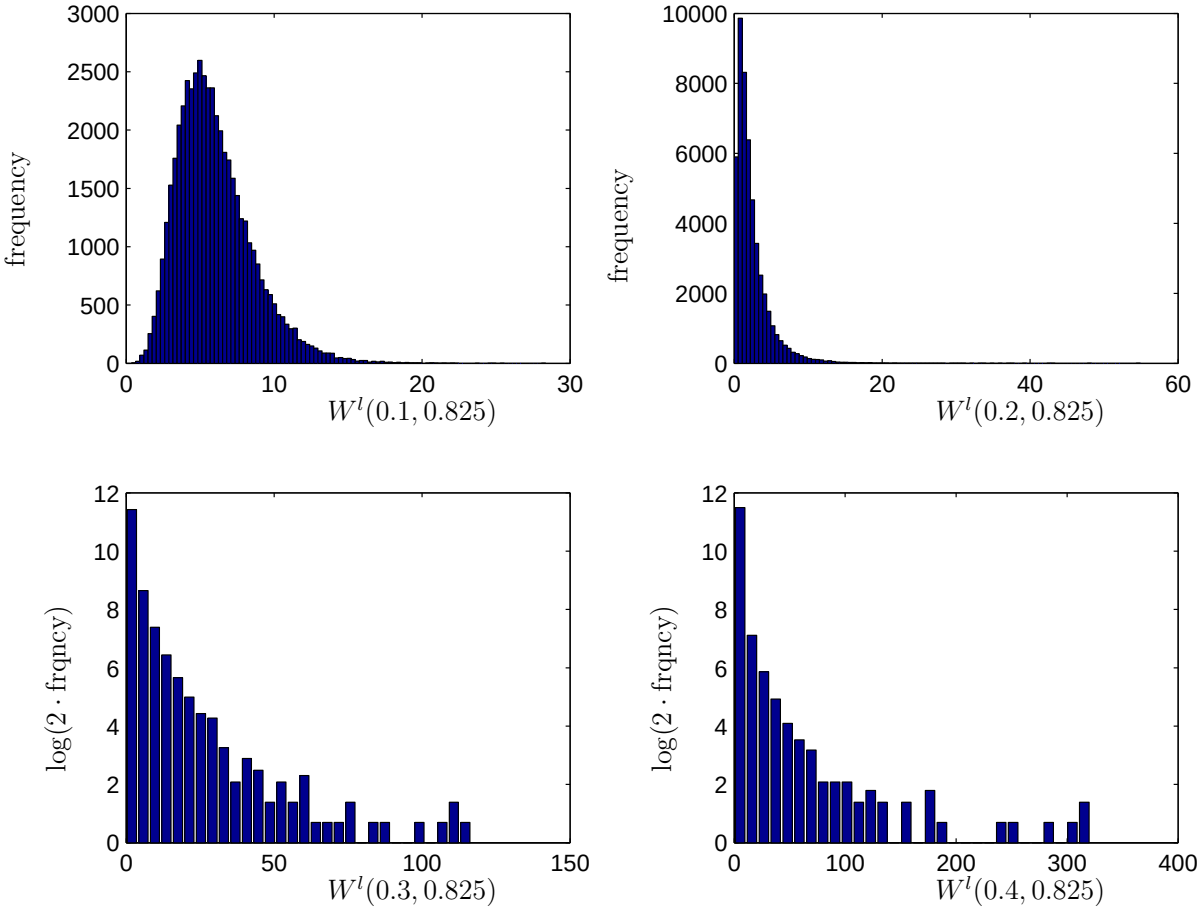

Figure 2: Histograms of 50000 samples of $W^l(t, q)$ for the coin flipping example with $m = 100$, $q = 0.825$, and for $t = 0.1, 0.2, 0.3$ and 0.4. In the last two histograms the ordinate is shown on a logarithmic scale

discussed example. The plots show the function

$$y \mapsto \log\Big(\mathbb{P}\big(W(t,q) > \mathrm{e}^y\big)\Big)$$

for $q = 0.825$ and $t = 0.3, 0.4$ respectively. It can be seen from the data that asymptotically the graph behaves like $-ay+b$, where $(a,b) = (2.22, 1.3)$ for $t = 0.3$ and $(a,b) = (1.6, 0.31)$ for $t = 0.4$. The value of a decreases with t.

For a reliable estimate of $VAR(W)$ via (5.42) the value of a would have to be substantially larger than 2, since for any distribution whose tail decay is governed by (5.43) we have $\mathbb{E}(W(t,q)) < +\infty$ if and only if $a > 1$ and $VAR(W(t,q)) < +\infty$ if and only if $a > 2$. Although for very large τ the tail decay of $W(t,q)$ is exponential, the fact that (5.43) holds for intermediate to large values of τ renders the variance estimation via (5.42) unreliable.

Therefore, useful information about the distribution of (5.42) *is not available, at least for reasonably small values of* l_0 *and reasonably large values of* t*.*

To further illustrate this point, Figure 3 also shows the histograms of 500 samples of $\widehat{VAR}_{100}\big(W(t,q)\big)$ for $q = 0.825$ and $t = 0.3, 0.4$ respectively, for the coin flipping example with $m = 100$. The ordinate of the second histogram, corresponding to $t = 0.4$, is reported on a logarithmic scale. Heavy tails of the distribution of $\widehat{VAR}_{100}\big(W(t,q)\big)$ are apparent because of the occurrence of massive outliers. The tails become lighter only very slowly with increasing l_0. For example, the tails of $\widehat{VAR}_{1000}(W(t,q))$ are heavier than those of the variable $\widehat{var}_{p,1000,1}$ which is defined below and whose histogram appears in Figure 4.

5.3.3 Avoiding the Pitfalls

In the previous paragraph we argued that the evaluating the empirical variance (5.42) is unreliable for certain values of (t,q). How can this problem be overcome?

On the one hand, one could impose an upper bound on t, depending on the value of m, so as to guarantee that (5.42) does not have major outliers. In fact, one can argue that in the typical range of t where (5.41) takes its optimum the distribution of (5.42) usually does not have too heavy tails. This observation forms the basis of a practical version of our method and implementations in Visual C++ undertaken in the recent MSc thesis [16]. A drawback of this approach is that certain assumptions about the distribution of (5.42) observed for moderately small values of m and l_0 can not be verified experimentally in reasonable time for the typical values of m and l_0 used in actual computations.

We are therefore going to pursue a different approach: recall that we are interested in the empirical variance (5.42) only because it is an unbiased estimator of $VAR\big(W(t,q)\big)$.

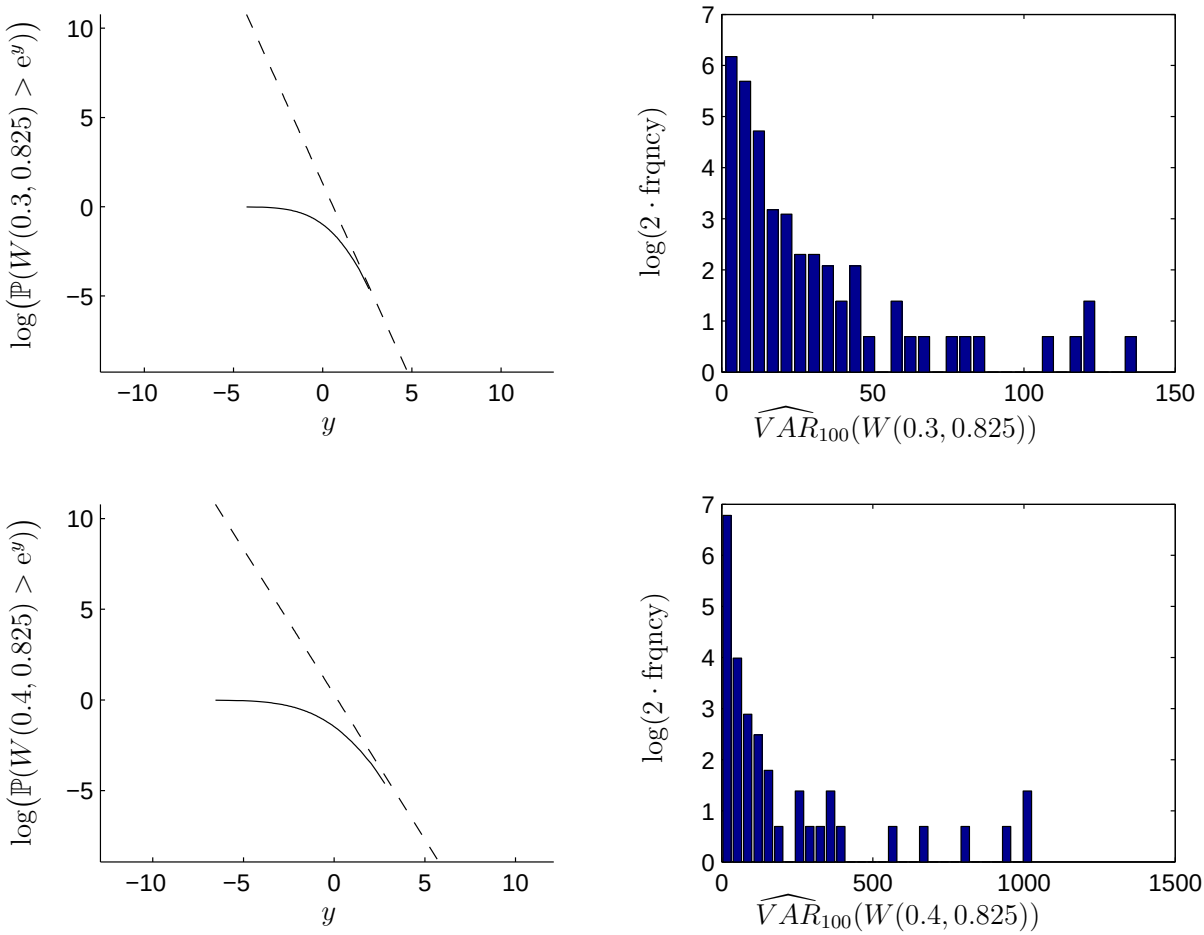

Figure 3: Empirical distribution tails of $W(t,q)$ are algebraic. Solid lines represent the function $\log\big(\mathbb{P}(W(t,q) > \mathrm{e}^y)\big)$, dashed lines the function $ay + b$. Histograms of 500 samples of $\widehat{VAR}_{100}(W(t,q))$ for $t = 0.3, 0.4$ respectively, plotted on a logarithmic scale on the ordinate.

But likewise, so is

$$\frac{1}{r}\sum_{j=1}^{r}\left(\Big(\frac{1}{s}\sum_{k=1}^{s}W^{l_{i,j,k}}(t,q)\Big)-\mathbb{E}[W(t,q)]\right)^2$$

for each $(i=1,\dots,p)$ when $W^{l_{i,j,k}}(t,q)$ $(i=1,\dots,p;j=1,\dots,r;k=1,\dots,s)$ are i.i.d. copies of $W(t,q)$. Choosing $l_0=prs$ large enough, $\mathbb{E}[W(t,q)]\simeq\frac{1}{prs}\sum_{i,j,k}W^{l_{i,j,k}}(t,q)$, so that

$$\widehat{var}^{\,i}_{p,r,s}(t,q):=\frac{1}{r-1}\sum_{j=1}^{r-1}\left(\Big(\frac{1}{s}\sum_{k=1}^{s}W^{l_{i,j,k}}(t,q)\Big)-\frac{1}{prs}\sum_{i,j,k}W^{l_{i,j,k}}(t,q)\right)^2 \tag{5.44}$$

is approximately an unbiased estimator of $VAR\big(W(t,q)\big)$.

Note that (5.44) is an improved version of the empirical variance of $\frac{1}{s}\sum_{k=1}^{s}W^{l_{i,j,k}}(t,q)$, $(j=1,\dots,r)$ for a fixed i, the only difference being that $\frac{1}{rs}\sum_{j,k}W^{l_{i,j,k}}$ has been replaced by $\frac{1}{prs}\sum_{i,j,k}W^{l_{i,j,k}}(t,q)$ which can be expected to have better converged to $\mathbb{E}[W(t,q)]$. This replacement achieves a slight lightening of the distribution tails, and it also introduces a small bias in the direction of overestimating, which we don't mind, because our aim is to derive an upper bound on $VAR(W(t,q))$.

The really powerful advantage of the new variance estimator is the fact that it is computed on the basis of the averaged data $\frac{1}{s}\sum_{k=1}^{s}W^{l_{i,j,k}}(t,q)$, which still have algebraic empirical decay but with a rate that becomes faster with increasing s. Indeed, Figure 4 shows that (5.44) has much lighter distribution tails than (5.42). In order to make the advantages of averaging apparent, we computed these histograms using the data underlying the first histogram of Figure 3. In the first row of Figure 4 we chose $(q,r,s)=(50,1000,1)$, that is, no averaging was applied. In the second row $(q,r,s)=(50,200,5)$, and in the third row $(q,r,s)=(50,10,100)$ was chosen. The plots in the left hand column show

$$\log\mathbb{P}\left(\frac{1}{s}\sum_{k=1}^{s}W^{l_{i,j,k}}(0.3,0.825)>\mathrm{e}^{y}\right)$$

as a function of y for $s=1,5,100$ respectively. Note that the asymptotes of the graphs have decreasing gradient with increasing s, corresponding to a faster algebraic decay of the empirical distribution tails of the averaged data. Not surprisingly, when the decay rate of the averaged data becomes sufficiently fast the distribution of (5.44) increasingly resembles a Gaussian: a Lilliefors test applied to the data of the third histogram does not reject the hypothesis that the data is Gaussian with a p-value of 17.9%. The second histogram shows that even averaging of only 5 independent copies of $W(t,q)$ achieves a remarkable decrease in the heaviness of distribution tails and renders the distribution much more symmetric.

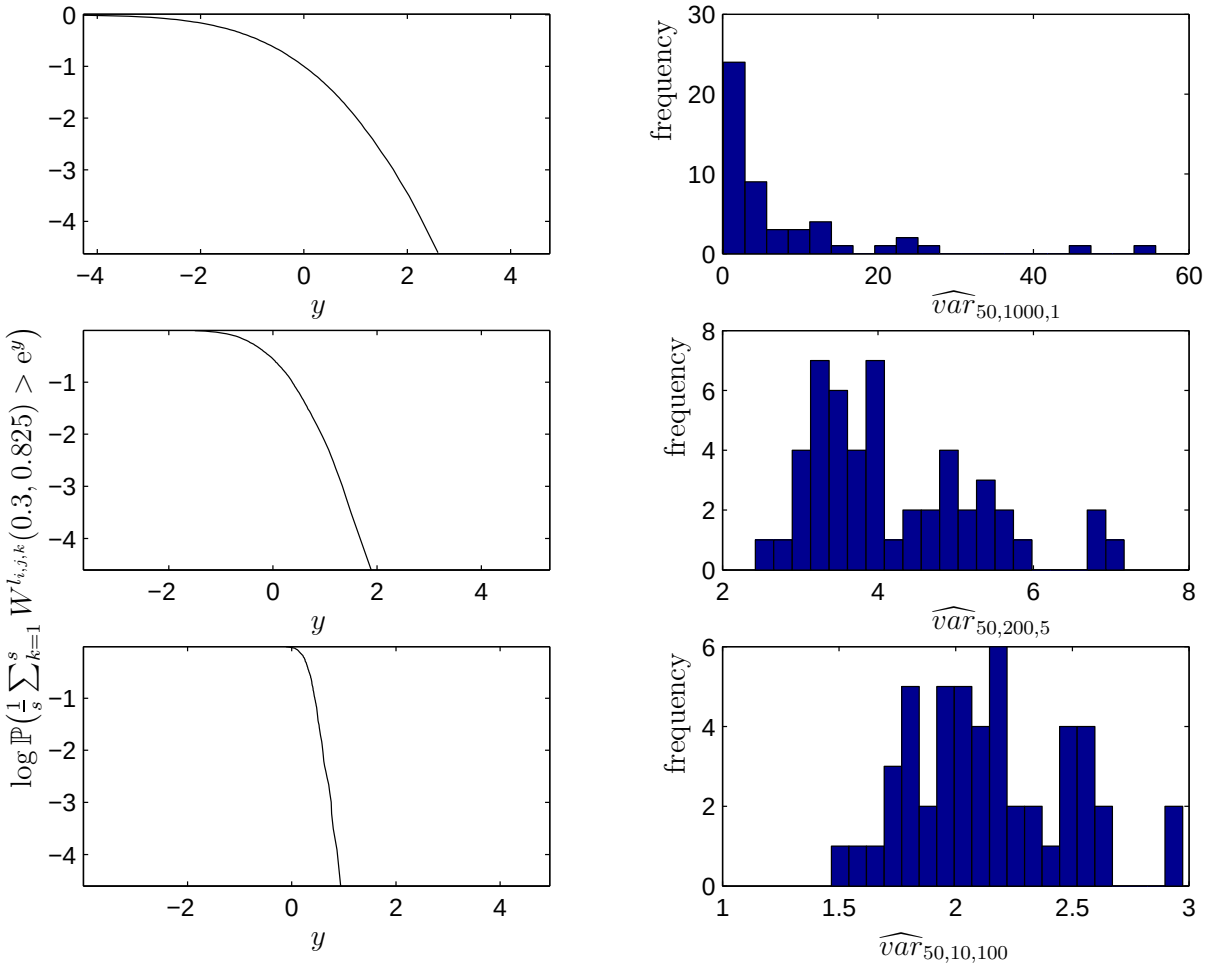

Figure 4: Algebraic decay rate of the tails of the empirical distribution tails of $W(t, q)$ becomes sharper with averaging. Variance estimates based on averaged data become increasingly Gaussian.

5.3.4 A Practical Algorithm

Note that if the distribution of a random variable X is perfectly symmetric, then the probability that 8 out of 10 independent copies $X_1, \ldots, X_{10}$ of X lie below $\mathbb{E}[X]$ equals

$$\mathbb{P}\big(|\{i : X_i \leq \mathbb{E}[X]\}| \geq 8\big) = 0.0107. \tag{5.45}$$

We will assume that if a Lilliefors test on the 5% level does not reject the hypothesis that $\widehat{var}_{10,r,s}$ is Gaussian, then the distribution is sufficiently symmetric for (5.45) to hold approximately for $X = \widehat{var}_{10,r,s}$. Since we know moreover that $\mathbb{E}\big[\widehat{var}_{10,r,s}\big] > VAR(W)$, we are confident that in this case

$$\mathbb{P}\Big(VAR\big(W(t,q)\big) > \widehat{var}^{[8]}_{10,r,s}(t,q)\Big) \leq 0.0107 = \eta \cdot (1-\Lambda), \tag{5.46}$$

with $\eta = 0.214$ and $\Lambda = 0.95$, and where $\widehat{var}^{[8]}_{10,r,s}$ denotes the 8-th order statistic of $\widehat{var}^{i}_{10,r,s}$ $(i = 1, \ldots, 10)$.

Our practical algorithm for finding an upper bound on the Chvàtal-Sankoff constant γ on the confidence level $\Lambda = 95\%$ is thus as follows:

1. For given input A and ξ, set $\Lambda = 95\%$, $\eta = 0.214$, $p = 10$ and choose m, r and s.
2. Generate the data vectors Z^l $(l = 1 \ldots, l_0 = p \cdot r \cdot s)$ using variant of the Wagner-Fischer algorithm.
3. For $t > 0$, $q \in (0,1)$, evaluate $\hat{v}(t,q) = \widehat{var}^{[8]}_{10,r,s}(t,q)$.
4. Determine $(\hat{t}, \hat{q})$ by solving the optimization problem (5.41).
5. If a Lilliefors test on the 5% level rejects the hypothesis that $\widehat{var}^{i}_{10,r,s}(\hat{t},\hat{q})$ $(i = 1, \ldots, 10)$ are Gaussian data, then increase s and/or r and return to Step 2. Otherwise accept $\hat{q}$ as an upper bound on γ on the 95% confidence level.

The last step provides a tool for automatically determining the number of simulations l_0 necessary for the results to be reliable on the 95% confidence level: it guarantees that the assumption on the symmetry of the distribution of $\widehat{var}10, r, s$ holds reasonably well at the optimal values of t and q. Of course, the method can be adapted to other values of p and Λ, but we found that our choice are reasonable values for the limited computing power of a desktop machine.

6 Implementation and Numerical Results

A straightforward adaptation of the Wagner-Fischer algorithm [24], which is based on dynamic programming, can compute the set of all m-matches contained in a pair of infinite random sequences (X, Y) in $O(m^2)$ time. A careful implementation which avoids computing unnecessary matrix entries achieves a practical complexity which is in effect closer to $O(m \log m)$. Moreover, the method can be implemented in such a way that only $O(m)$ storage of information is needed at any time point during a run of the algorithm. This is important because implementations based on $O(m^2)$ storage quickly spend most of the execution time moving information between different hierarchies of memory. The nontrivial constraint in the optimization problem (5.41) was strictly convex in all examples we attempted. Therefore, it is easy to find the global minimizer using standard software tools. We chose the Sequential Quadratic Programming solver of the Matlab Optimization Toolbox which could solve all examples to a precision of 10^{-8} within a few iterations. The Lilliefors test is implemented in standard software packages. We chose to use the Matlab Statistics Toolbox in which the test can be performed using a simple Matlab command.

The method was implemented in Matlab 6.1 and experiments were run on a SunBlade 100 workstation. Our aim was numerical accuracy and reliability rather than speed, and there remains considerable room for optimizing the code from the latter perspective, for example by removing multiple loops and by working with sparse matrix data structures.

In our experiments we considered the LCS problems in which A has $|A| = 2, \ldots, 4$ characters and where ξ is the uniform measure. Each of the experiments reported in Table 6 took a few days to complete. The value of $\hat{q}$ did not change significantly after a few hundred simulations, but we continued simulating until the variance data $\widehat{var}^i_{10,r,s}(\hat{t}, \hat{q})$ did not reject the Lilliefors test on the $P = 5\%$ level and hence was sufficiently symmetric. In all four experiments we chose $m = 1000$, $p = 10$ and $\Lambda = 0.95$, that is, $\hat{q}$ is an upper bound on γ at the 95% confidence level. The p-value P of the Lilliefors test and the number s of independent copies of Z used in averaging the raw data are listed in the last two columns. For comparison we also list the best deterministic lower and upper bounds for these examples, denoted by DLB and DUP respectively, which were derived by Dančik-Paterson [15, 22], as well as the best known probabilistic lower and upper bounds at the 95% confidence level, denoted by ALB and AUP respectively, which were derived by Alexander [2] on the basis of two simulations of $E[L_50000]$. Finally, we list the best known probabilistic lower bounds BLB *without confidence guarantee* which were obtained by Baeza-Yates, Gavalda, Navarro and Scheihing [10] on the basis of ten simulations of $E[L_{100000}]$.

Although for much larger m roundoff errors might play a significant role, such effects

$\|A\|$	DLB	ALB	BLB	$\hat{\mathbf{q}}$	DUP	AUP	P	s	l_0
2	0.7739	0.8079	0.8118	**0.8182**	0.8376	0.8607	0.0675	400	8000
3	0.6338	–	0.7172	**0.7235**	0.7658	–	>0.2	400	12000
4	0.5528	–	0.6537	**0.6601**	0.7082	–	>0.2	200	8000

Table 1: New upper bounds $\hat{q}$ on the 95% confidence level are computed with $m = 1000$. A comparison with BLB shows that $\hat{q}$ approximates the true value of γ to about $5 \cdot 10^{-3}$.

are minimal for $m = 1000$. For example, in the case $|A| = 2$ it is easy to see that the worst-case round-off error for the nontrivial constraint function

$$c(t,q) := \frac{1}{l_0}\sum_{l=1}^{l_0} W^l(t,q) + \sqrt{\frac{\hat{v}(t,q)}{(1-\eta)(1-\Lambda)l_0}} - 1$$

of (5.41) can be bounded by 10^{-9}. On the other hand, at the optimal values $(\hat{t},\hat{q})$ one finds that $\left|\frac{\partial}{\partial q}c(\hat{t},\hat{q})\right| > 10^{-2}$. Therefore, the backward error $\Delta\hat{q}$ satisfies $10^2 \cdot \Delta\hat{q} < 10^{-9}$, that is $\Delta\hat{q} < 10^{-11}$. However, since $\hat{q}$ approximates γ only to about $5 \cdot 10^{-3}$, we have $\hat{q} - \gamma > 10^{-4} \gg \Delta\hat{q}$. This shows that rounding errors neither wrongly indicate that $\hat{q}$ is an upper bound on γ nor interfere with its approximating quality.

References

[1] David Aldous and Persi Diaconis. Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem. *Bull. Amer. Math. Soc. (N.S.)*, 36(4):413–432, 1999.

[2] Kenneth S. Alexander. The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Probab.*, 4(4):1074–1082, 1994.

[3] A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, editors. *Combinatorial pattern matching*, volume 684 of *Lecture Notes in Computer Science*, Berlin, 1993. Springer-Verlag.

[4] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.*, 17(1):9–25, 1989.

[5] R. Arratia, L. Gordon, and M.S. Waterman. The Erdős-Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.*, 18(2):539–570, 1990.

[6] R. Arratia and M.S. Waterman. The Erdős-Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.*, 17(3):1152–1169, 1989.

[7] Richard Arratia, Louis Gordon, and Michael Waterman. An extreme value theory for sequence matching. *Ann. Statist.*, 14(3):971–993, 1986.

[8] Richard Arratia and Michael S. Waterman. A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, 4(1):200–225, 1994.

[9] K. Azuma. Weighted sums of certain dependent random variables. *Tohuku Math. J.*, 19:357–367, 1967.

[10] R.A. Baeza-Yates, R. Gavaldà, G. Navarro, and R. Scheihing. Bounding the expected length of longest common subsequences and forests. *Theory Comput. Syst.*, 32(4):435–452, 1999.

[11] Jinho Baik, Percy Deift, and Kurt Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999.

[12] Renato Capocelli, Alfredo De Santis, and Ugo Vaccaro, editors. *Sequences. II.* Springer-Verlag, New York, 1993. Methods in communication, security, and computer science, Papers from the workshop held in Positano, June 17–21, 1991.

[13] Renato M. Capocelli, editor. *Sequences.* Springer-Verlag, New York, 1990. Combinatorics, compression, security, and transmission, Papers from the workshop held in Naples and Positano, June 6–11, 1988.

[14] Václáv Chvatal and David Sankoff. Longest common subsequences of two random sequences. *J. Appl. Probability*, 12:306–315, 1975.

[15] Vlado Dančík and Mike Paterson. Upper bounds for the expected length of a longest common subsequence of two binary sequences. *Random Structures Algorithms*, 6(4):449–458, 1995.

[16] Quentin Decouvelaere. *Upper Bounds for the LCS Problem.* MSc Thesis, Oxford University Computing Laboratory, 2003.

[17] Joseph G. Deken. Some limit results for longest common subsequences. *Discrete Math.*, 26(1):17–31, 1979.

[18] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statis. Assoc.*, 58:13–30, 1963.

[19] Krengel. *Ergodic Theorems.* W. de Gruyter, Berlin-New York, 1985.

[20] Joseph B. Kruskal. An overview of sequence comparison: time warps, string edits, and macromolecules. *SIAM Rev.*, 25(2):201–237, 1983.

[21] Claudia Neuhauser. A Poisson approximation for sequence comparisons with insertions and deletions. *Ann. Statist.*, 22(3):1603–1629, 1994.

[22] Mike Paterson and Vlado Dančík. Longest common subsequences. In *Mathematical foundations of computer science 1994 (Košice, 1994)*, volume 841 of *Lecture Notes in Comput. Sci.*, pages 127–142. Springer, Berlin, 1994.

[23] David Sankoff and Joseph B. Kruskal, editors. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison.* Addison-Wesley Publishing Company Advanced Book Program, Reading, MA, 1983.

[24] Robert A. Wagner and Michael J. Fischer. The sring-to-string correction problem. *J. Ass. Comp. Mach.*, 21:168–173, 1974.

[25] M. Waterman. *Introduction to Computational Biology.* Chapman & Hall, 1995.

[26] Michael S. Waterman. General methods of sequence comparison. *Bull. Math. Biol.*, 46(4):473–500, 1984.

[27] Michael S. Waterman. Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B*, 344:383–390, 1994.

[28] M.S. Waterman and M. Vingron. Sequence comparison significance and poisson approximation. *Statistical Science*, 9(3):367–381, 1994.