Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications xx (xxxx) xxx–xxx

stochastic processes and their applications

www.elsevier.com/locate/spa

Lower bounds for moments of global scores of pairwise Markov chains

Jüri Lember[a,*], Heinrich Matzinger[b], Joonas Sova[a], Fabio Zucca[c]

[a] *Institute of Mathematics and Statistics, University of Tartu, J. Liiv 2, 50409 Tartu, Estonia*
[b] *School of Mathematics, Georgia Tech, Atlanta, (GA), 30332, USA*
[c] *Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy*

Received 17 August 2016; received in revised form 4 August 2017; accepted 10 August 2017
Available online xxxx

Abstract

Let $X_1, \dots$ and $Y_1, \dots$ be random sequences taking values in a finite set $\mathbb{A}$. We consider a similarity score $L_n := L(X_1, \dots, X_n; Y_1, \dots, Y_n)$ that measures the homology of words $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$. A typical example is the length of the longest common subsequence. We study the order of moment $E|L_n - EL_n|^r$ in the case where the two-dimensional process $(X_1, Y_1), (X_2, Y_2), \dots$ is a Markov chain on $\mathbb{A} \times \mathbb{A}$. This general model involves independent Markov chains, hidden Markov models, Markov switching models and many more. Our main result establishes a condition that guarantees that $E|L_n - EL_n|^r \asymp n^{\frac{r}{2}}$. We also perform simulations indicating the validity of the condition.

MSC: 60K35; 41A25; 60C05

Keywords: Random sequence comparison; Longest common sequence; Fluctuations; Waterman conjecture

* Corresponding author.
E-mail addresses: juri.lember@ut.ee (J. Lember), matzi@math.getech.edu (H. Matzinger), joonas.sova@ut.ee (J. Sova), fabio.zucca@polimi.it (F. Zucca).

http://dx.doi.org/10.1016/j.spa.2017.08.009

1. Introduction

1.1. Sequence comparison setting

Throughout this paper $X = (X_1, X_2, \dots, X_n)$ and $Y = (Y_1, Y_2, \dots, Y_n)$ are two random strings, usually referred as sequences, so that every random variable X_i and Y_i takes values on a finite alphabet $\mathbb{A}$. Since the sequences X and Y are not necessarily independent nor identically distributed, it is convenient to consider the two-dimensional sequence $Z = ((X_1, Y_1), \dots, (X_n, Y_n))$. The sample space of Z will be denoted by $\mathcal{Z}_n$. Clearly $\mathcal{Z}_n \subseteq (\mathbb{A} \times \mathbb{A})^n$ but, depending on the model, the inclusion can be strict.

The problem of measuring the similarity of X and Y is central in many areas of applications including computational molecular biology [10,16,40,42,46] and computational linguistics [33,34,36,37]. In this paper, we consider a general scoring scheme, where $S : \mathbb{A} \times \mathbb{A} \to \mathbb{R}^+$ is a *pairwise scoring function* that assigns a score to each couple of letters from $\mathbb{A}$. An *alignment* is a pair (ρ, τ) where $\rho = (\rho_1, \rho_2, \dots, \rho_k)$ and $\tau = (\tau_1, \tau_2, \dots, \tau_k)$ are two increasing sequences of natural numbers, i.e. $1 \le \rho_1 < \rho_2 < \cdots < \rho_k \le n$ and $1 \le \tau_1 < \tau_2 < \cdots < \tau_k \le n$. The integer k is the number of aligned letters, $n - k$ is the number of non-aligned letters. Given the pairwise scoring function S the score of the alignment (ρ, τ) when aligning X and Y is defined by

$$U_{(\rho,\tau)}(X, Y) := \sum_{i=1}^{k} S(X_{\rho_i}, Y_{\tau_i}) + \delta(n - k),$$

where $\delta \in \mathbb{R}$ is another scoring parameter. Typically $\delta \le 0$ so that many non-aligned letters in the alignment reduce the score. If $\delta \le 0$, then its absolute value $|\delta|$ is often called the *gap penalty*. Given S and δ, the optimal alignment score of X and Y is defined to be

$$L_n := L(X, Y) = L(Z) := \max_{(\rho,\tau)} U_{(\rho,\tau)}(X, Y), \tag{1.1}$$

where the maximum above is taken over all possible alignments. Sometimes, when we talk about a *string comparison model*, we refer to the study of L_n for given sequences X and Y, score function S and gap penalty δ. It is important to note that for any constant gap price $\delta \in \mathbb{R}$, changing the value of one of the $2n$ random variables $X_1, \dots, X_n, Y_1, \dots, Y_n$ changes the value of L_n by at most Δ, where

$$\Delta := \max_{u,v,w \in \mathbb{A}} \Big(|S(u, v) - S(u, w)| \vee |S(u, v) - S(w, v)| \Big). \tag{1.2}$$

When $\delta = 0$ and the scoring function assigns one to every pair of similar letters and zero to all other pairs, i.e.

$$S(a, b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b \end{cases}, \tag{1.3}$$

then $L(Z)$ is just the maximal number of aligned letters, also called the length of the *longest common subsequence* (abbreviated by LCS) of X and Y. In this article, to distinguish the length of LCS from another scoring schemes, we shall denote it via $\ell_n := \ell(Z) = \ell(X, Y)$. In other words $\ell(Z)$ is the maximal k so that there exists an alignment (ρ, τ) such that $X_{\rho_i} = Y_{\tau_i}$, $i = 1, \dots, k$. Note that the optimal alignment (ρ, τ) as well as the longest common subsequence $X_{\rho_1}, \dots, X_{\rho_k}$ is not typically unique. The length of LCS is probably the most important and the most studied measure of global similarity between strings.

Please cite this article in press as: J. Lember, et al., Lower bounds for moments of global scores of pairwise Markov chains, Stochastic Processes and their Applications (2017), http://dx.doi.org/10.1016/j.spa.2017.08.009.