J. Appl. Prob. 25, 106–119 (1988)
Printed in Israel

STOCHASTIC SCRABBLE: LARGE DEVIATIONS FOR SEQUENCES WITH SCORES

RICHARD ARRATIA,*†§
PRICILLA MORRIS* AND
MICHAEL S. WATERMAN,*†† *University of Southern California*

Abstract

A derivation of a law of large numbers for the highest-scoring matching subsequence is given. Let X_k, Y_k be i.i.d. $\boldsymbol{q} = (q(i))_{i \in S}$ letters from a finite alphabet S and $\boldsymbol{v} = (v(i))_{i \in S}$ be a sequence of non-negative real numbers assigned to the letters of S. Using a scoring system similar to that of the game Scrabble, the score of a word $w = i_1 \cdots i_m$ is defined to be $V(w) = v(i_1) + \cdots + v(i_m)$. Let V_n denote the value of the highest-scoring matching contiguous subsequence between $X_1X_2 \cdots X_n$ and $Y_1Y_2 \cdots Y_n$. In this paper, we show that $V_n/K \log(n) \to 1$ a.s. where $K \equiv K(\boldsymbol{q}, \boldsymbol{v})$. The method employed here involves 'stuttering' the letters to construct a Markov chain and applying previous results for the length of the longest matching subsequence. An explicit form for $\boldsymbol{\beta} \in \Pr(S)$, where $\beta(i)$ denotes the proportion of letter i found in the highest-scoring word, is given. A similar treatment for Markov chains is also included.

Implicit in these results is a large-deviation result for the additive functional, $H \equiv \Sigma_{n<\tau} v(X_n)$, for a Markov chain stopped at the hitting time τ of some state. We give this large deviation result explicitly, for Markov chains in discrete time and in continuous time.

LARGE DEVIATIONS; ADDITIVE FUNCTIONALS; DNA SEQUENCE MATCHING; MARKOV CHAINS

1. Introduction

This paper gives an extension of the results of Arratia and Waterman [2] on the length of the longest matching subsequence. Here we consider the case where a matching subsequence, or word w, is scored according to the letters appearing in the match. The scoring system is similar to that of the board game Scrabble; hence the name stochastic Scrabble.

The motivation for this problem came from the comparison of DNA sequences which are sometimes modeled as sequences of i.i.d. letters or as letters of a Markov chain, from a four-letter alphabet: A = adenine, C = cytosine, G = guanine, and T = thymine. Given

Received 26 November 1985; revision received 13 October 1986.

Postal address for all authors: Department of Mathematics, University of Southern California, Los Angeles, CA 90089–1113, USA.

Research supported by * the System Development Foundation, † the National Science Foundation, § the Institute of Mathematics and its Applications and †† the National Institutes of Health.

some *a priori* knowledge of which matching letters signify a higher degree of similarity, scoring techniques provide a measure which can be used to compare different DNA sequences. Frequently, runs of letters of a certain type are related to biological signals. For example, segments of sequence rich in As and Ts are frequently involved in regulation of the transcription of DNA encoding proteins. In comparing two regions of DNA which might be of this type, finding matching A/T segments is of great interest. The topic of sequence comparison is reviewed in [7] where these and other topics are discussed. Obviously it is important to study the probability distribution of the maximum score for the case of random sequences.

Let $X_1, X_2, \cdots$ and $Y_1, Y_2, \cdots$ be i.i.d. letters over a finite alphabet S with common distribution $\boldsymbol{q}$. Let $\boldsymbol{v}$ be a non-negative vector assigning scores to the letters of S. The score of a word $w = i_1 \cdots i_m$ is $V(w) = v(i_1) + \cdots + v(i_m)$ so that the value of the highest-scoring matching subsequence is

$$\begin{aligned} V \equiv V_n \equiv \max\{ V(w)\colon & \text{ there exists } 1 \leqq m \leqq n \text{ such that} \\ & w = X_{i+1} \cdots X_{i+m} = Y_{j+1} \cdots Y_{j+m} \text{ for some } 0 \leqq i, j \leqq n-m \}. \end{aligned}$$

If $v(i) = 1$ for all $i \in S$, then $V_n = M_n \equiv$ length of the longest matching subsequence (i.e. allowing shifts), given by

$$M \equiv M_n \equiv \max\{m : X_{i+k} = Y_{j+k}, \text{ for } k = 1 \text{ to } m \text{ for some } 0 \leqq i, j \leqq n-m\}.$$

In [2], it is shown that for i.i.d. letters, M_n follows an Erdös–Rényi law. That is,

$$\boldsymbol{P}\left(\lim_{n \to \infty} M_n / \log_{1/p}(n) = K \right) = 1,$$

where $p = \boldsymbol{P}(X_1 = Y_1)$ and $k = 2$. Without shifts, this is analogous to the length of the longest head-run R_n, for which Erdös and Rényi proved $K = 1$ (see [4] and [6]). [2] also treated the case where the X_i, Y_i are independent Markov chains (irreducible and aperiodic), with transition probabilities $[p_{ij}]$. The result is

$$(1) \qquad \boldsymbol{P}((M_n - 2\log_{1/p}(n))/\log_{1/p}\log_{1/p}(n) \in (-4-\varepsilon, 1+\varepsilon) \text{ eventually}) = 1,$$

where p is the largest eigenvalue of the substochastic matrix $P = [(p_{ij})^2]$.

In Section 3, we show that

$$\boldsymbol{P}((V_n - 2\log_{1/p}(n))/\log_{1/p}\log_{1/p}(n) \in (-4-\varepsilon, 1+\varepsilon) \text{ eventually}) = 1,$$

where p is the largest solution to the equation:

$$(2) \qquad 1 - (q_1)^2 \lambda^{-v(1)} - (q_2)^2 \lambda^{-v(2)} - \cdots - (q_d)^2 \lambda^{-v(d)} = 0.$$

Our method is to transform the problem from V_n to M_n by 'stuttering' the letters into Markov chains and applying (1). In Section 4, we show that a slight modification in the form of condition (2) occurs when stuttering is applied to already existing Markov chains; however, the result remains basically unchanged. In both cases a discussion on the composition of the highest scoring word is given.

The case where the sequences are i.i.d. but with different distributions or different lengths, is handled in [1] using large-deviation methods. Results from [1] are used when deriving $\boldsymbol{\beta} \in \mathrm{Pr}(S)$, where $\beta(i)$ denotes the proportion of letter i found in the highest-scoring word.

2. Transformation to a Markov process

Let $X_1, X_2, \cdots$ and $Y_1, Y_2, \cdots$ be i.i.d. random sequences over a finite alphabet $S = \{1, 2, \cdots, d\}$ with common distribution $\boldsymbol{q} = (q_1, \cdots, q_d)$. Let $\boldsymbol{v} = (v(1), \cdots, v(d))$ be a sequence of positive integers assigned to the letters of S.

Consider one of the sequences, say $\{X_i\}$, and define a new sequence $\{\tilde{X}_i\}$ by duplicating each letter in the sequence according to its value assigned by $\boldsymbol{v}$. Thus,

$$X_1 X_2 \cdots X_n = i_1 i_2 \cdots i_n$$

gives

$$\tilde{X}_1 \tilde{X}_2 \cdots \tilde{X}_{\tilde{N}(n)} = \underbrace{i_1 i_1 \cdots i_1}_{v(i_1)\text{ times}} \underbrace{i_2 i_2 \cdots i_2}_{v(i_2)\text{ times}} \cdots \underbrace{i_n i_n \cdots i_n}_{v(i_n)\text{ times}} .$$

The new sequence $\{\tilde{X}_i\}$ is now a Markov chain, over the alphabet

$$S' = \{(1, 1), \cdots, (1, v(1)), \cdots, (d, 1), \cdots, (d, v(d))\},$$

with random length $\tilde{N}(n) = v(i_1) + \cdots + v(i_n)$ which is governed by $\boldsymbol{q}$ and $\boldsymbol{v}$. Relabeling the alphabet S' with the integers $1, 2, \cdots, |S'|$, we get the transition probabilities over the relabeled alphabet $\tilde{S} = \{1, 2, \cdots, \sigma(d)\}$ (where $\sigma(d) = |S'| = \Sigma_{1 \leq k \leq d} v(k)$):

$$\begin{aligned}
&P_{i,i+1} = 1 && \text{for } i \neq \sigma(1), \cdots, \sigma(d) \\
&P_{\sigma(i),\sigma(j)+1} = q_{j+1} && \text{for } i = 1, \cdots, d; j = 1, \cdots, d-1 \\
&P_{\sigma(i),1} = q_1 && \text{for } i = 1, \cdots, d \\
&P_{i,j} = 0 && \text{otherwise,}
\end{aligned} \tag{3}$$

where $\sigma(i) = \Sigma_{1 \leq k \leq i} v(k)$.

For example, consider the i.i.d. random sequence $\{X_i\}$ over $S = \{a, b\}$ with scores $\boldsymbol{v} = (2, 3)$ such that $\boldsymbol{P}(X_1 = a) = \frac{2}{3}$ and $\boldsymbol{P}(X_1 = b) = \frac{1}{3}$. The transition probability matrix for $\tilde{X}_1$ is

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{2}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \frac{2}{3} & 0 & \frac{1}{3} & 0 & 0 \end{bmatrix}$$

Consider a second sequence $\{Y_i\}$, and perform the same transformation to define $\{\tilde{Y}_i\}$. The score of a word $w = i_1 i_2 \cdots i_m$ is $V(w) = \Sigma_{1 \leq k \leq d} v(i_k)$, which is identical to the length of the transformed word $\tilde{w}$. Therefore, the highest-scoring matching word between $\{X_i\}$ and $\{Y_i\}$ is identical with the length of the longest matching word between the transformed sequences $\{\tilde{X}_i\}$ and $\{\tilde{Y}_i\}$. Thus the problem is changed from V_n, $\{X_i\}$, $\{Y_i\}$ to one concerning $M_{\tilde{N}(n)}$, $\{\tilde{X}_i\}$, $\{\tilde{Y}_i\}$. Formally stated, if $\{X_i\}$ and $\{Y_i\}$ are i.i.d. random sequences over S and $\{\tilde{X}_i\}$ and $\{\tilde{Y}_i\}$ are independent Markov chains with transition probabilities and state space given as in (3), then $V_n = M_{\tilde{N}(n)}$, where V_n is taken with respect to $\{X_i\}$, $\{Y_i\}$ and $M_{\tilde{N}(n)}$ is taken with respect to $\{\tilde{X}_i\}$, $\{\tilde{Y}_i\}$.

Note that the lengths of the new Markov sequences are random and not necessarily of the same length. To be precise, $N^x(n)$ and $N^y(n)$ are each the sum of independent random variables, given by

$$N^x(n) = \sum_{1 \leq k \leq n} \xi_k \quad \text{and} \quad N^y(n) = \sum_{1 \leq k \leq n} \varphi_k,$$

where $\xi_k = v(i)$ iff $X_k = i$ and $\varphi_k = v(i)$ iff $Y_k = i$. Let $c = \Sigma_{i \in S}\, v(i) q_i$. Since $\xi_1, \xi_2, \cdots$ are i.i.d. with $\boldsymbol{P}(\xi_1 = v(i)) = \boldsymbol{P}(X_1 = i) = q_i$, an application of the strong law of large numbers shows that

$$\lim_{n \to \infty} N^x(n)/n = c \quad \text{almost surely.} \tag{4}$$

The same statement may be made for $N^y(n)$.

In [2], extensions of (1) are stated for the case of matching between sequences of different lengths. However, by using (4), it is still possible to apply (1) in its simplest form. Consider the stuttered sequences $\{\tilde{X}_i\}$, $\{\tilde{Y}_i\}$, with $V_n = M_{\tilde{N}(n)}$. Set $N_1(n) = \min\{N^x(n), N^y(n)\}$ and $N_2(n) = \max\{N^x(n), N^y(n)\}$. By (4) it follows that

$$\lim_{n \to \infty} N_1(n)/n = \lim_{n \to \infty} N_2(n)/n = c \quad \text{almost surely.} \tag{5}$$

Truncating the (infinite) sequences at $N_1(n)$ and $N_2(n)$ and using the monotonicity of M_k we get that, for any n,

$$M_{N_1(n)} \leqq \boldsymbol{V_n} \leqq M_{N_2(n)}.$$

Thus it suffices to consider the random variable $M_{N_1(n)}$ (matching between sequences of equal lengths) since by (5), all pertinent statements about $M_{N_1(n)}$ will also hold for $M_{N_2(n)}$. For convenience we define $\tilde{N}(n)$ to be $N_1(n)$.

3. Scoring with i.i.d. letters

Through the transformation just described we are now able to formulate our results via an analysis of $M_{\tilde{N}(n)}$ with respect to the Markov processes (3) governed by $[p_{ij}]$ combined with the law of large numbers (1).

Inspection of the transition probability matrix for $\tilde{X}$ and $\tilde{Y}$ shows that it is irreducible, but not necessarily aperiodic. In fact, $\gcd\{v(1), \cdots, v(d)\} > 1 \Leftrightarrow P$ is not aperiodic. However, without loss of generality, we may assume aperiodicity since the weights can always be scaled so that $\gcd\{v(1), \cdots, v(d)\} = 1$. Furthermore, it can be checked (by direct calculation), that the left and right eigenvectors corresponding to the eigenvalue λ, for $P = [(p_{ij})^2]$ are:

$$
\boldsymbol{l}=\begin{bmatrix} l(1,1)\\ l(1,2)\\ \cdot\\ \cdot\\ l(1,\nu(1))\\ \cdot\\ \cdot\\ \cdot\\ l(d,1)\\ l(d,2)\\ \cdot\\ \cdot\\ l(d,\nu(d) \end{bmatrix}^{t}=\begin{bmatrix} (q_1)^2\lambda^{-1}\\ (q_1)^2\lambda^{-2}\\ \cdots\\ \cdots\\ (q_1)^2\lambda^{-\nu(1)}\\ \cdots\\ \cdots\\ \cdots\\ (q_d)^2\lambda^{-1}\\ (q_d)^2\lambda^{-2}\\ \cdots\\ \cdots\\ (q_d)^2\lambda^{-\nu(d)} \end{bmatrix}^{t}
$$

$$
\boldsymbol{r}=\begin{bmatrix} r(1,1)\\ r(1,2)\\ \cdot\\ \cdot\\ r(1,\nu(1))\\ \cdot\\ \cdot\\ \cdot\\ r(d,1)\\ r(d,2)\\ \cdot\\ \cdot\\ r(d,\nu(d)) \end{bmatrix}=\begin{bmatrix} \lambda^{-\nu(1)}\\ \lambda^{-\nu(1)+1}\\ \cdots\\ \cdots\\ \lambda^{-1}\\ \cdots\\ \cdots\\ \cdots\\ \lambda^{-\nu(d)}\\ \lambda^{-\nu(d)+1}\\ \cdots\\ \cdots\\ \lambda^{-1} \end{bmatrix}\cdot\lambda \tag{6}
$$

where λ is a root of the equation

$$1-(q_1)^2\lambda^{-\nu(1)}-\cdots-(q_d)^2\lambda^{-\nu(d)}=0. \tag{7}$$

Note that the equation in λ has a real root $p\in(0,1)$. Frobenius' theorem for positive matrices ensures that the largest real root of (7) is the largest eigenvalue for $P=[(p_{ij})^2]$ [4].

Theorem 1. Let $\boldsymbol{\nu}\in[0,\infty)^d$, for all $\varepsilon>0$,

$$\boldsymbol{P}((V_n-2\log_{1/p}(n))/\log_{1/p}\log_{1/p}(n)\in(-4-\varepsilon,1+\varepsilon)\text{ eventually})=1,$$

where $p\in(0,1)$, is the largest root of the equation

$$f(\lambda)\equiv 1-(q_1)^2\lambda^{-\nu(1)}-\cdots-(q_d)^2\lambda^{-\nu(d)}=0.$$

Proof. Suppose all $\nu(i)$ are positive integers. By the transformation described in Section 2, we have that $V_n=M_{\tilde{N}(n)}$, where $\tilde{N}(n)/n\to c$ by (5), and M_k is the longest matching subsequence between independent Markov chains which satisfy the conditions of (1). Then, for fixed $\varepsilon>0$,

$$\boldsymbol{P}((M_k - 2\log_{1/p}(k))/\log_{1/p}\log_{1/p}(k) \in (-4-\varepsilon, 1+\varepsilon) \text{ eventually}) = 1,$$

where $p \in (0, 1)$ satisfies (7) and $p > |\lambda|$ for all $\lambda \neq p$ which also satisfy (7). By (5) we have that

$$\boldsymbol{P}(\tilde{N}(n) \in (nc/2, 2nc) \text{ eventually}) = 1.$$

Using the monotonicity of M_k this implies

$$\boldsymbol{P}(M_{[nc/2]} \leqq M_{\tilde{N}(n)} \leqq M_{[2nc]} \text{ eventually}) = 1,$$

where $[x]$ denotes the integral part of x. The deterministic error bounds for all ω, as $n \to \infty$

$$|(M_{[nc/2]} - 2\log_{1/p}(n) + 2\log_{1/p}[c/2])/\log_{1/p}(\log_{1/p}(n) + \log_{1/p}[c/2])$$
$$- (M_{[nc/2]} - 2\log_{1/p}(n))/\log_{1/p}\log_{1/p}(n)| \to 0,$$

combined with the almost sure statement above yields that for all $\varepsilon > 0$,

$$\boldsymbol{P}((M_{[nc/2]} - 2\log_{1/p}(n))/\log_{1/p}\log_{1/p}(n) \in (-4-\varepsilon, 1+\varepsilon) \text{ eventually}) = 1.$$

Similarly one can show that $M_{[2nc]}$ also falls within the desired range. This takes care of the case where all $v(i)$ are positive integers.

For $v(i) \in [0, \infty)$, approximate the vector $\boldsymbol{v}$ by a sequence of rational-valued vectors $\boldsymbol{v}_t$ scaled by t to contain all integer components.

For each i, let $v_t(i) = \max\{[tv(i)], 1\}$ and $\boldsymbol{v}_t = (v_t(1), \cdots, v_t(d))$. Let $V_{t,n}$ = max score under the weighting $\boldsymbol{v}_t$. Use the fact that $v_t(i)/t \to v(i)$ as $t \to \infty$ to obtain: for all n, ω

$$\frac{V_n}{V_{t,n}/t} \to 1 \quad \text{as } t \to \infty.$$

Apply the argument above to $V_{t,n}$ and replace $\eta(t) = (\lambda)^t$ in

$$f(\lambda; t) = 1 - (q_1)^2\lambda^{-v_t(1)} - \cdots - (q_d)^2\lambda^{-v_t(d)} = 0$$

to get, for all $\varepsilon > 0$,

$$\boldsymbol{P}((V_{t,n} - 2\log_{1/\eta(t)}(n))/\log_{1/\eta(t)}\log_{1/\eta(t)}(n) \in (-4-\varepsilon, 1+\varepsilon) \text{ eventually}) = 1,$$

where $\eta(t)$ is the largest solution of

$$f(\eta; t) = 1 - (q_1)^2\eta^{-v_t(1)/t} - \cdots - (q_d)^2\eta^{-v_t(d)/t} = 0.$$

Since $\sup_{0 \leqq \eta \leqq 1}|f(\eta, t) - f(\eta)| \to 0$ as $t \to \infty$, we get $\eta(t) \to p$ which is the largest real solution for $f(\eta)$ as stated in the theorem.

We now turn to a result on the composition of the highest scoring word. Let $\Pr(S')$ denote the space of probability measures on S'. Our method is to use a result from [1] to obtain $\boldsymbol{\pi} \in \Pr(S')$ such that $\pi(i)$ = expected proportion of times letter i occurs in a long matching word. Converting back to the original alphabet S is then relatively straightforward.

In [1] it is shown, for matching between two independent Markov processes, with M_n denoting the length of the longest match between $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$,

$$1 = \boldsymbol{P}\left(0 = \lim_{n\to\infty} \max_{a\in S'} \sup \left\{ \left| \pi(a) - (1/m) \sum_{t=1 \text{ to } m} \boldsymbol{l}(X_{i+t} = a) \right| : \right.\right.$$

$$\left.\left. X_{i+t} = Y_{j+t}, \text{ for some } 0 \leqq i, j \leqq n - m \text{ and } t = 1, \cdots, m, \text{ where } m = M_n \right\}\right),$$

where $(\pi(i)) = (r(i)l(i))$, and $r(i)$ and $l(i)$ are the components of the right and left eigenvectors corresponding to the principal eigenvalue p. Informally, this says that the proportion of letters in every longest matching word tends almost surely to $(\pi(i)) = (r(i)l(i))$.

Thus, for $v(i) \in \{\mathbb{Z}^+\}^d$, we use $\boldsymbol{l}$ and $\boldsymbol{r}$ from (6) (ignoring the common factor $1/p$), to get: $\pi(i, 1) = \pi(i, 2) = \cdots = \pi(i, v(i)) = (q_i)^2 p^{-v(i)}$ for each $i \in S$. To obtain a vector $\boldsymbol{\beta} = (\beta(1), \cdots, \beta(d))$ over the original alphabet, set $\beta(i) = \pi(i, 1)$ and normalize so that $\boldsymbol{\beta} \in \Pr(S)$. Then $\boldsymbol{\beta}$ has the form

$$\boldsymbol{\beta} = (\pi(1, 1), \cdots, \pi(d, 1))/c_0$$

and $\beta(i)$ = proportion of letter i found in the highest-scoring word. Here, the normalizing constant $c_0 = 1$ by Equation (7). For $v(i) \in [0, \infty)$, use the approximation method as in the proof of Theorem 1. Then $\pi_t(i, 1) \to \pi(i, 1) = (q_i)^2 p^{-v(i)}$ for all $i \in S$, and $\eta(t) \to p$, where p is the largest solution to

$$(8) \qquad 1 - (q_1)^2 \lambda^{-v(1)} - \cdots - (q_d)^2 \lambda^{-v(d)} = 0.$$

Theorem 2. The proportion of letter i found in every highest-scoring word converges almost surely to $\beta(i)$, where $\boldsymbol{\beta}$ has the form

$$\boldsymbol{\beta} = ((q_1)^2 p^{-v(1)}, \cdots, (q_d)^2 p^{-v(d)})$$

and p is the largest solution to Equation (8).

The form of $\boldsymbol{\beta}$ given in Theorem 2 can also be verified by the methods of [1].

Example 1. Suppose $\{X_i\}$ and $\{Y_i\}$ are two sequences of fair-coin tosses with a match of heads being given a weight of 1, and a match of tails a weight of 2 (i.e. $S = \{1, 2\}$, $\boldsymbol{v} = (1, 2)$ and $\boldsymbol{q}$ is uniform). Then,

$$P = [(p_{ij})^2] = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix}, \quad 0 = 1 - (\tfrac{1}{4})p^{-2} - (\tfrac{1}{4})p^{-1},$$

and $\boldsymbol{\pi} = (p^{-1}, p^{-2}, p^{-2})/4(p^{-1} + 2p^{-2})$. Using Theorem 2: $\boldsymbol{\beta} = (p, 1)/(p + 1)$. Since $p = (1 + (17)^{1/2})/8$ we get: $\boldsymbol{\beta} \approx (0.3716265, 0.6283735)$. Changing the scoring to $\boldsymbol{v} = (0, 1)$ we get: $\boldsymbol{\beta} = (0.25, 0.75)$.

Example 2. Consider the following generalization of the example above.

Let $S = \{1, 2, \ldots, 2n\}$ and $\boldsymbol{v} = (1, 2, 1, \cdots, 2)$, so that $v(\text{odd}) = 1$ and $v(\text{even}) = 2$, with $\boldsymbol{q}$ uniform over S. Then the equation to be solved is

$$1-(1/4n)p^{-1}-(1/4n)p^{-2}=0$$

which has the solution $p=(1+(1+16n)^{1/2})/8n$. For $n=1$ we get Example 1. For arbitrary n, $\boldsymbol{\beta}$ has the form

$$\boldsymbol{\beta}=(p^{-1},p^{-2},p^{-1},\cdots,p^{-2})/4n^2.$$

Notice that:

$$\sum\boldsymbol{\beta}(\text{odd})=1/4np=2/(1+(1+16n)^{1/2})\rightarrow 0\quad\text{as } n\rightarrow\infty.$$

Thus the size of the alphabet has a strong effect on the composition of the highest-scoring word. Compare this with a generalization of the second part of Example 1. Using the same set-up, change the scoring vector to $\boldsymbol{\nu}=(0,1,0,\cdots,1)$. The new equation

$$1-(1/4n)p^{-0}-(1/4n)p^{-1}=0,$$

is satisfied by $p=1/(4n-1)$. For this $\boldsymbol{\nu}$, we find

$$\boldsymbol{\beta}=(1,1/p,1,\cdots,1/p)/(n+n/p),$$

and $\Sigma\,\boldsymbol{\beta}(\text{odd})=1/4n\rightarrow 0$ as $n\rightarrow\infty$. Again we find that the appearance of odd letters decays with increases in the size of the alphabet. However, in this case the rate at which appearances of odd letters decays is much faster.

4. Scoring with sequences of independent Markov chains

The discussion for Markov chains closely follows that of the i.i.d. case. Let X_i be a Markov chain over a finite alphabet S with transition probabilities $[p_{ij}]$. If $\boldsymbol{\nu}=(\nu(1),\cdots,\nu(d))$ has positive integer components, we can apply the same transformation as in the i.i.d. case to get $\tilde{P}$ given by:

$$\begin{aligned}
&\tilde{P}_{i,i+1}=1 && \text{for } i\neq\sigma(1),\cdots,\sigma(d)\\
&\tilde{P}_{\sigma(i),\sigma(j)+1}=(p_{i,j+1}) && \text{for } i=1,\cdots,d;\ j=1,\cdots,d-1\\
&\tilde{P}_{\sigma(i),1}=(p_{i,1}) && \text{for } i=1,\cdots,d\\
&\tilde{P}_{i,j}=0 && \text{otherwise}
\end{aligned}$$

where $\sigma(i)=\Sigma_{1\leq k\leq i}\,\nu(k)$.

Theorem 3. Let X_i, Y_i be independent Markov chains over $S=\{1,\cdots,d\}$ with transition probabilities $[p_{ij}]$, with $p_{ij}>0$ for all i,j. Let $\nu(i)\in[0,\infty)$ for all $i\in S$. Then, for all $\varepsilon>0$

$$\boldsymbol{P}((V_n-2\log_{1/p}(n))/\log_{1/p}\log_{1/p}(n)\in(-4-\varepsilon,1+\varepsilon)\text{ eventually})=1,$$

where p is the largest root of $\det(P-\lambda^V)=0$, with $P=[(p_{ij})^2]$, $\lambda^V=[\delta_{ij}\lambda^{\nu(i)}]$.

Proof. Let $Q=[(\tilde{p}_{ij})^2]$. The right eigenvector for Q must have the form

$$\boldsymbol{r}=(r(1),\lambda r(1),\cdots,\lambda^{\nu(1)-1}r(1),\cdots,r(d),\lambda r(d),\cdots,\lambda^{\nu(d)-1}r(d))'.$$

If $Qr = \lambda r$, the system of equations

$$\begin{array}{llr}
((p_{11})^2 - \lambda^{v(1)})r(1) + (p_{12})^2 r(2) + \cdots + & & (p_{1d})^2 r(d) = 0 \\
(p_{21})^2 r(1) + ((p_{22})^2 - \lambda^{v(2)})r(2) + \cdots + & & (p_{2d})^2 r(d) = 0 \\
\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \\
\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \\
(p_{d1})^2 r(1) + (p_{d2})^2 r(2) + \cdots + & & ((p_{dd})^2 - \lambda^{v(d)})r(d) = 0
\end{array}$$

is satisfied by λ. Equivalently,

$$\det(P - \lambda^V) = 0, \quad \text{where } P = [(p_{ij})^2] \text{ and } \lambda^V = [\delta_{ij}\lambda^{v(i)}]. \tag{9}$$

We claim that λ is an eigenvalue for Q, iff it is a root of (9). Let $P(t) = D + t(P - D)$ where $D = \text{diag}(P)$. Then $f_t(\lambda) = \det(P(t) - \lambda^V)$ and its roots $\lambda_i(t)$ are continuous in t. If $t = 0$ then $P(t) = D$, and $f_t(\lambda)$ has $n = v(1) + \cdots + v(d)$ roots — one for each eigenvalue of $Q(t)$. As t varies continuously from 0 to 1, each $\lambda_i(t)$ traces a continuous arc in C so that $\lambda_i(1) = \lambda_i$ (i.e. the eigenvalues of Q), multiplicities included. Thus the roots of $f_1(\lambda) = \det(P - \lambda^V) = 0$ represent all the eigenvalues of Q. The theory of positive matrices now implies that the largest root of (9) is the largest eigenvalue for Q. Applying the rational approximation procedure for $v(i) \in [0, \infty)$, as in Theorem 1, completes the proof.

By the argument used in the proof we know that r can be found by using Cramer's rule on the matrix $[P - (p)^V]$, where p is the largest solution to (9). Thus we have that

$$r(1) = M(d, 1), r(2) = M(d, 2), \cdots, r(d-1) = -M(d, d-1), r(d) = M(d, d),$$

where $M(i, j) =$ minor of the (i, j)th entry of the matrix $[P - (p)^V]$, and

$$\boldsymbol{r} = \begin{bmatrix} r(1,1) \\ r(1,2) \\ \cdot \\ \cdot \\ r(1, v(1)) \\ \cdot \\ \cdot \\ \cdot \\ r(d,1) \\ r(d,2) \\ \cdot \\ \cdot \\ r(d, v(d)) \end{bmatrix} = (\pm) \begin{bmatrix} r(1) \\ \lambda r(1) \\ \cdots \\ \cdots \\ \lambda^{v(1)-1} r(1) \\ \cdots \\ \cdots \\ \cdots \\ r(d) \\ \lambda r(d) \\ \cdots \\ \cdots \\ \lambda^{v(d)-1} r(d) \end{bmatrix}$$

Similarly one can show

$$\boldsymbol{l}=\begin{bmatrix} l(1,1)\\ l(1,2)\\ \cdot\\ \cdot\\ l(1,v(1))\\ \cdot\\ \cdot\\ \cdot\\ l(d,1)\\ l(d,2)\\ \cdot\\ \cdot\\ l(d,v(d)) \end{bmatrix}^{t} = (\pm)\begin{bmatrix} \lambda^{v(1)-1}\boldsymbol{l}(1)\\ \lambda^{v(1)-2}\boldsymbol{l}(1)\\ \cdots\\ \cdots\\ \boldsymbol{l}(1)\\ \cdots\\ \cdots\\ \cdots\\ \lambda^{v(d)-1}\boldsymbol{l}(d)\\ \lambda^{v(d)-2}\boldsymbol{l}(d)\\ \cdots\\ \cdots\\ \boldsymbol{l}(d) \end{bmatrix}^{t}$$

where $\boldsymbol{l}(1)=M_t(d,1)$, $\boldsymbol{l}(2)=M_t(d,2),\cdots,\boldsymbol{l}(d-1)=-M_t(d,d-1)$, $\boldsymbol{l}(d)=M_t(d,d)$ and $M_t(i,j)$ are the minors of the (i,j)th entry of the matrix $[P^t-(p)^V]$. Then $\boldsymbol{\pi}$ can be given by $\pi(i,1)=\cdots=\pi(i,v(i))=p^{v(i)}r(i)l(i)/c$ for all $i\in S$, where c is a normalizing constant.

Theorem 4. The proportion of letter i found in every highest-scoring word converges almost surely to $\beta(i)$, where $\boldsymbol{\beta}$ has the form

$$\boldsymbol{\beta}=(p^{v(1)}r(1)l(1),\cdots,p^{v(d)}r(d)l(d))/c,$$

and p is the largest solution to (9), and c is a normalizing constant.

Example. Let $S=\{1,2\}$, $\boldsymbol{v}=(1,2)$ and $p_{ij}=0.75$ if $i=j$ and $p_{ij}=0.25$ otherwise.

$$P=\begin{bmatrix}(0.75)^2 & (0.25)^2\\ (0.25)^2 & (0.75)^2\end{bmatrix} \quad\text{and}\quad (P-\lambda^V)=\begin{bmatrix}(0.75)^2-\lambda & (0.25)^2\\ (0.25)^2 & (0.75)^2-\lambda^2\end{bmatrix}=0.$$

Solving we find $p\approx 0.76288$, so that

$$\boldsymbol{r}\approx(\pm)\begin{bmatrix}-0.0625\\ -0.20038\\ -0.152865\end{bmatrix} \quad\text{and}\quad \boldsymbol{l}\approx(\pm)\begin{bmatrix}-0.0625\\ -0.152865\\ -0.20038\end{bmatrix}^{t}$$

Combined with Theorem 4, this gives: $\boldsymbol{\beta}\approx(0.113996, 0.8869039)$. Changing the scoring to $\boldsymbol{v}=(0,1)$ we get: $\boldsymbol{\beta}\approx(0.0190598, 0.9809402)$.

5. Large deviations

Theorems 1 and 3 above give a precise version of the statement '$V_n\approx 2\log_{1/p}(n)$.' These theorems can be viewed as a synthesis of two results: first, a large-deviation result giving the value of p; and second, a correlation bound to establish the factor of 2. The large-deviation result includes the case where the scoring function v takes both positive and negative values and may be of independent interest, so we present it below as Theorem 5. A continuous-time version of this is given as Theorem 5′. The role of a correlation bound in Theorems 1 and 3 is not apparent in this paper, because our proof refers back to [2], where the correlation bound was exploited.

In the absence of a correlation bound, the large deviation result easily determines the asymptotic growth rate of M_n, but only to within a factor of 2. We carry this out in Theorem 6. Further results about the correlation between matches at different pairs of positions, with applications to the longest match-run, appear in [2], [3], and [8].

Theorem 5. Let $X_0, X_1, X_2, \cdots$ be a Markov chain on $S = \{0, 1, \cdots, d\}$ with transition probabilities $[p_{ij}]$ such that all states $i, j \neq 0$ communicate, 0 is accessible, and $P_{00} = 1$. Let $v : S \to \boldsymbol{R}$, let τ be the hitting time for 0, and let

$$H = \sum_{0 \leqq n < \tau} v(X_n).$$

Assume that for all $t > 0$, $\boldsymbol{P}(H > t) > 0$. Then

$$\lim_{t \to \infty} t^{-1} \log \boldsymbol{P}(H \geqq t) = \log p,$$

where p is a root in (0, 1) of $\det(P - \lambda^V) = 0$, with $P = [(p_{ij})]$ and $\lambda^V = [\delta_{ij}\lambda^{v(i)}]$, $i, j = 1$ to d.

Proof. First we show the existence of $\lambda_{cr} \in (0, 1)$, satisfying

$$\log(\lambda_{cr}) \equiv \lim_{t \to \infty} t^{-1} \log \boldsymbol{P}(H \geqq t \mid X_0 = i) \in (-\infty, 0) \quad \text{for } i = 1 \text{ to } d.$$

To do this, let $c \equiv \max_i v(i)$ and $1/b = \min_i P(H \geqq 0) \mid X_0 = i$. For $s, t > 0$ and $i \neq 0$, a stopping-time argument shows that

$$\boldsymbol{P}(H \geqq s + t + c \mid X_0 = i) \leqq b\boldsymbol{P}(H \geqq s \mid X_0 = i) \max_j \boldsymbol{P}(H \geqq t \mid X_0 = j).$$

Thus if we let $g(t) = \max_i b\boldsymbol{P}(H \geqq t - c \mid X_0 = i)$, then $g(t + s) \leqq g(t)g(s)$, for $s, t > c$ so $\lim_{t \to \infty} t^{-1} \log[g(t)]$ exists. Irreducibility implies that for $i = 1$ to d,

$$\lim_{t \to \infty} t^{-1} \log \boldsymbol{P}(H \geqq t \mid X_0 = i) = \lim_{t \to \infty} t^{-1} \log[g(t)]. < 0$$

The condition that, for all t, $\boldsymbol{P}(H \geqq t) \geqq 0$ is equivalent to the condition that there exists a 'possible cycle of positive score', i.e. for some $i(0), i(1), \cdots, i(k) = i(0) \in \{1, 2, \cdots, d\}$, $p_{i(j-1), i(j)} > 0$ for $j = 1$ to k, and $\Sigma_{j=1 \text{ to } k}\, v(i(j)) > 0$. This shows that $\lim_{t \to \infty} t^{-1} \log[g(t)] > -\infty$, so that $\lambda_{cr} \in (0, 1)$.

Second, let $f_i(\lambda) \equiv \boldsymbol{E}(\lambda^{-H} \mid X_0 = i) \in (0, \infty]$, for $i = 0$ to d and $\lambda \in (0, 1)$. By considering the value of j of X_1, we get the backward equations

$$f_i(\lambda) = \lambda^{-v(i)} \sum_{j = 0 \text{ to } d} p_{ij} f_j(\lambda)$$

for $i = 1$ to d, and $f_0(\lambda) = 1$. Rewrite these equations, for $\lambda > \lambda_{cr}$, as

$$(P - \lambda^V) f(\lambda) = -(p_{i0})_{i = 1 \text{ to } d},$$

with the matrix $P - \lambda^V$ and the vector $\boldsymbol{f}(\lambda)$ indexed by $i = 1$ to d. Clearly $\lambda_{cr} = \inf\{\lambda \in (0, 1] : f_i(\lambda) < \infty\}$ and $f_i(\lambda) \to \infty$ as $\lambda \to \lambda_{cr}^+$, for $i = 1$ to d. Irreducibility implies that the ratios $f_i(\lambda)/f_1(\lambda)$, for $i = 1$ to d, are bounded away from 0 and ∞ uniformly for $\lambda \in (\lambda_{cr}, 1]$. By compactness, there is a sequence $\lambda_n \downarrow \lambda_{cr}$ for which the limit $g_i \equiv \lim_{n\to\infty} f_i(\lambda_n)/f_1(\lambda_n) \in (0, \infty)$ all exist, for $i = 1$ to d. Thus, for $\lambda = \lambda_{cr}$, $(P - \lambda^V)\boldsymbol{g} = 0$, so that $\det(P - \lambda^V) = 0$.

Theorem 5′. Let $(X_t, t \geqq 0)$ be a continuous-time Markov process on $S = \{0, 1, \cdots, d\}$ with generator $Q = [q_{ij}]$, such that all states $i, j \neq 0$ communicate, 0 is accessible, and $q_{00} = 0$. Let $v : S \to \boldsymbol{R}$, with $v(i) > 0$ for at least one i, let τ be the hitting time for 0, and let

$$H = \int_0^\tau v(X_s)ds.$$

Then,

$$\lim_{t\to\infty} t^{-1} \log \boldsymbol{P}(H \geqq t) = -r$$

where $r \in (0, \infty)$ is a root of $\det(Q + xV) = 0$ and $Q + xV = [q_{ij} + \delta_{ij} xv\,(i)]_{i,j=1\text{ to } d}$.

Proof. The proof is essentially the same as the proof of Theorem 5, so we present the differences. Let $f_i(x) = \boldsymbol{E}(\exp(xH) \mid X_0 = i) \in (0, \infty]$, for $x \in (0, \infty)$. Note that $r \equiv -\lim_{t\to\infty} t^{-1} \log \boldsymbol{P}(H \geqq t)$ satisfies: for $i = 1, 2, \cdots, d$, $r = \sup\{x \in (0, \infty) : f_i(x) < \infty\}$ and $\lim_{x\to r-} f_i(x) = \infty$, while $f_0(x) = 1$ for all x. We get the backward equations by considering the time σ of the first jump away from i, say to state j. Thus, for $i = 1$ to d,

$$\begin{aligned} f_i(x) &= \sum_{j=0 \text{ to } d} \boldsymbol{E}(\exp(\sigma x v(i))\, f_j(x);\, X_\sigma = j) \\ &= \sum_{j=0 \text{ to } d,\, j\neq i} -q_{ij}/(xv(i) + q_{ii})\, f_j(x) \end{aligned}$$

if $xv(i) < -q_{ii}$; otherwise $f_i(x) = \infty$. Notice that $r \leqq \min_i(-q_{ii}/v(i))$, which corresponds to the possibility that H is large just because the process X remains a long time in its initial state. Rewrite these equations as: for $i = 1$ to d

$$\sum_{j=1 \text{ to } d} q_{ij} f_j(x) + xv(i)\, f_i(x) = -q_{i0},$$

or in matrix form:

$$(Q + xV)\,\boldsymbol{f} = -(q_{i0})_{i=1 \text{ to } d}.$$

Taking a limit as $x \uparrow r$ we get $(Q + rV)\boldsymbol{g} = 0$, where $\boldsymbol{g}$ is a vector with entries $g_i \in (0, \infty)$, $i = 1$ to d.

Notice that Theorem 5 can be applied to the scoring function $-\boldsymbol{v}$ to get the rate for large negative deviations of V, namely: if $\boldsymbol{P}(H < -t) > 0$ for all t, then $\lim_{t\to\infty} t^{-1} \log \boldsymbol{P}(H \geqq t) = -\log p^-$, where $p^- \in (1, \infty)$ is a root of $\det(P - \lambda^V) = 0$.

Similarly, if $v(i) < 0$ for some i, then Theorem 5′ can be applied to $-\boldsymbol{v}$ to prove: $\lim_{t\to\infty} t^{-1} \log \boldsymbol{P}(H < -t) = r^{-}$, where $r^{-} \in (-\infty, 0)$ is a root of $\det(Q + xV) = 0$.

Theorem 6. Let X_i, Y_i be independent Markov chains over $S = \{1, 2, \cdots, d\}$ with irreducible aperiodic transition probabilities $[p_{ij}]$. Let $v(i) \in (-\infty, \infty)$ and assume that for all t, there exists n such that $\boldsymbol{P}(V_n > t) > 0$. Then for all $\varepsilon > 0$,

$$\boldsymbol{P}(V_n/\log_{1/p}(n) \in [1-\varepsilon, 2+\varepsilon] \text{ eventually}) = 1,$$

where p is a root of $\det(P - \lambda^V) = 0$, with $P = [(p_{ij})^2]$, $\lambda^V = [\delta_{ij}\lambda^{v(i)}]$.

Proof. Fix $\varepsilon \in (0, \frac{1}{4})$. First we prove the upper bound. Using Theorem 5, for all sufficiently large t,

$$\boldsymbol{P}\left(\sum_{0 \leqq n < \tau} v(X_n) > t\right) < p^{t(1-\varepsilon)},$$

where $\tau = \inf\{n \geqq 0 : X_n \neq Y_n\}$. By decomposing the event $\{V_n \geqq t\}$ according to the indices $i, j \leqq n$ at which the high-scoring match appears, we get $\boldsymbol{P}(V_n \geqq t) \leqq n^2 p^{t(1-\varepsilon)}$. Using $t = 2(1 + 2\varepsilon)\log_{1/p}(n)$, this says

$$\boldsymbol{P}(V_n/\log_{1/p}(n) > 2 + 4\varepsilon) \leqq n^2 p^{t(1-\varepsilon)} = n^2 n^{-2(1+2\varepsilon)(1-\varepsilon)} < n^{-\varepsilon},$$

for sufficiently large n. Using the Borel–Cantelli lemma along a skeleton of times such as $n_k = 2^k$, it follows that

$$\boldsymbol{P}(V_n/\log_{1/p}(n) < 2 + 4\varepsilon \text{ eventually}) = 1.$$

Now we prove the lower bound. By Theorem 5, and using the hypothesis that each Markov chain is irreducible and aperiodic, there exists an integer k, depending on ε, such that

$$\boldsymbol{P}\left(\sum_{t \leqq n < \tau(t)} v(X_n) > t, \text{ and } \tau < kt \;\middle|\; X_0 = i,\, Y_0 = j\right) > p^{t(1+\varepsilon)},$$

for all $i, j \in S$ and all sufficiently large t, with $\tau(t) = \min\{n \geqq t : X_n \neq Y_n\}$. By considering only matches which begin at X_i and Y_i, where $i = j = lkt + t$ for some integer l, we see that for all positive integers m, and all sufficiently large t,

$$\boldsymbol{P}(V_{mkt} \leqq t \mid X_0 = i,\, Y_0 = j) < (1 - p^{t(1+\varepsilon)})^m < \exp(-mp^{t(1+\varepsilon)}), \quad \text{for all } i, j \in S.$$

Using $t = [(1 - 3\varepsilon)\log_{1/p}(n)]$ and $m = [n/kt]$, we have $p^t \geqq n^{-(1-3\varepsilon)}$ and

$$\boldsymbol{P}(V_n/\log_{1/p}(n) \leqq 1 - 3\varepsilon) \leqq \boldsymbol{P}(V_{mkt} \leqq t) < \exp\{-mp^{t(1+\varepsilon)}\}$$

$$< \exp\{-mn^{-(1-3\varepsilon)(1+\varepsilon)}\} < \exp(-n^{2\varepsilon}/kt) < \exp(-n^{\varepsilon}),$$

for all sufficiently large n. Using the Borel–Cantelli lemma, it follows that $\boldsymbol{P}(V_n \log_{1/p}(n) > 1 - \varepsilon \text{ eventually}) = 1$.

Acknowledgements

The authors would like to thank Louis Gordon for several very helpful discussions, and S. R. S. Varadhan for showing us the proof of Theorem 5 presented here.

References

[1] Arratia, R. and Waterman, M. S. (1985) Critical phenomena in sequence matching. *Ann. Prob.* **13**, 1236–1249.

[2] Arratia, R. and Waterman, M. S. (1985) An Erdös–Rényi law with shifts. *Adv. Math.* **55**, 13–23.

[3] Arratia, R., Gordon, L., and Waterman, M. S. (1986) An extreme value theory for sequence matching. *Ann. Statist.* **14**, 971–993.

[4] Erdös, P. and Rényi, A. (1970) On a new law of large numbers. *J. Anal. Math.* **22**, 103–111.

[5] Karlin, S. and Taylor, H. M. (1970) *A First Course in Stochastic Processes*, 2nd edn. Academic Press, New York.

[6] Rényi, A. (1970) *Probability Theory*, Akademia Kiado, Budapest.

[7] Waterman, M. S. (1984) General methods of sequence comparison. *Bull. Math. Biol.* **46**, 473–500.

Reference added in proof

[8] Arratia, R., Goldstein, L. and Gordon, L. (1988) Two moments suffice for Poisson approximations: the Chen–Stein method. *Ann. Prob.* To appear.