J. Appl. Prob. **56,** 558–573 (2019)
*doi:*10.1017/jpr.2019.26

ON THE RATE OF CONVERGENCE FOR THE LENGTH OF THE LONGEST COMMON SUBSEQUENCES IN HIDDEN MARKOV MODELS

C. HOUDRÉ* ** AND
G. KERCHEV* *** *Georgia Institute of Technology*

Abstract

Let $(X, Y) = (X_n, Y_n)_{n\geq 1}$ be the output process generated by a hidden chain $Z = (Z_n)_{n\geq 1}$, where Z is a finite-state, aperiodic, time homogeneous, and irreducible Markov chain. Let LC_n be the length of the longest common subsequences of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. Under a mixing hypothesis, a rate of convergence result is obtained for $\mathrm{E}[LC_n]/n$.

Keywords: Longest common subsequence; rate of convergence; hidden Markov model; Hoeffding inequality; mixing condition

2010 Mathematics Subject Classification: Primary 60C05; 05A05
Secondary 60F10

1. Introduction

Longest common subsequences are often a key measure of similarity between two strings of letters. For two finite sequences $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ taking values in a finite alphabet $\mathcal{A}$, the object of study is $LCS(X_1, \ldots, X_n; Y_1, \ldots, Y_m)$, the length of the longest common subsequences of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, which is abbreviated as LC_n when $n = m$. Clearly, LC_n is the largest k such that there exist $1 \leq i_1 < \cdots < i_k \leq n$ and $1 \leq j_1 < \cdots < j_k \leq n$ with

$$X_{i_s} = Y_{j_s}, \quad \text{for all } s = 1, 2, 3, \ldots, k.$$

For two independent words sampled independently and uniformly at random from the alphabet, Chvátal and Sankoff [6] proved that $\lim_{n\to\infty} \mathrm{E}[LC_n]/n = \gamma^*$ and provided upper and lower bounds on γ^*. This was followed by Alexander [1], who obtained, for indpendent and identically distributed (i.i.d.) draws, the following generic rate of convergence result:

$$n\gamma^* - C\sqrt{n \log n} \leq \mathrm{E}[LC_n] \leq n\gamma^*, \tag{1.1}$$

where $C > 0$ is an absolute constant.

From a practical point of view the independence assumptions, both between words and also among draws, has to be relaxed as they are often lacking. One such instance is in the field of computational biology where we compare similarities between two biological sequences. In particular, alignments of those sequences need to be qualified as occurring by chance or

Received 1 May 2018; revision received 12 February 2019.
* Postal address: School of Mathematics, Georgia Institute of Technology, 686 Cherry Street Atlanta, GA 30332-0160, USA.
** Email address: houdre@math.gatech.edu
*** Email address: kerchev@math.gatech.edu

because of a structural relation. One way to generate alignments is with a hidden Markov model (HMM). The states of the hidden chain account for a match between two elements in X and Y or for an alignment of an element with a gap. Given X and Y we can find the most probable alignment using the Viterbi algorithm. This model is particularly useful when the similarity between X and Y is weak. In this case standard methods for pairwise alignment often fail to identify the correct alignment or test for its significance. With a hidden Markov model we can evaluate the total probability that X and Y are aligned by summing over all alignments, and this sum can be efficiently computed with the Forward algorithm. For more information the reader is referred to [9, Chapter 4].

There are very few results on the asymptotics of the longest common subsequences in a model exhibiting dependence properties. A rare instance is due to Steele [18], who showed the convergence of $\mathrm{E}[LC_n]/n$ when (X, Y) is a random sequence for which there is a stationary ergodic coupling, e.g. an irreducible, aperiodic, positive recurrent Markov chain. The present paper studies the longest common subsequences for strings exhibiting a different Markov relation; namely, we study the case when (X, Y) is emitted by a latent Markov chain Z, i.e. when $(Z, (X, Y))$ is a hidden Markov model. Note that this framework includes the special case when (Z, X) and (Z', Y) are hidden Markov models, with the same parameters, while Z and Z' are independent. In this setting, mean convergence is quickly proved in Section 2. Then, the main contribution is a rate of convergence result, obtained in Section 3, which recovers, in particular, (1.1).

Throughout this manuscript the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is assumed to be rich enough to consider all the random variables being studied.

2. Mean convergence

Recall that a hidden Markov model (Z, V) consists of a Markov chain $Z = (Z_n)_{n\geq 1}$ which emits the observed variables $V = (V_n)_{n\geq 1}$. The possible states in Z are each associated with a distribution on the values of V. In other words, the observation V is a mixture model where the choice of the mixture component for each observation depends on the component of the previous observation. The mixture components are given by the sequence Z. Note also that given Z, V is a Markov chain. For such a model the first easy result asserts the mean convergence of LC_n.

Proposition 2.1. *Let Z be an aperiodic, irreducible, time-homogeneous, finite-state-space Markov chain. Let μ, P, and π be respectively the initial distribution, transition matrix, and stationary distribution of Z. Let each Z_n, $n \geq 1$, generate a pair (X_n, Y_n) according to a distribution associated with the state of Z_n, i.e. let $(Z, (X, Y))$ be a hidden Markov model, where $X = (X_n)_{n\geq 1}$ and $Y = (Y_n)_{n\geq 1}$. Further, for all $i \geq 1$ and $j \geq 1$, let X_i and Y_j take their values in the common finite alphabet $\mathcal{A}$ and let there exist $a \in \mathcal{A}$ such that $\mathrm{P}(X_i = Y_j = a) > 0$ for some $i \geq 1$ and $j \geq 1$. Then*

$$\lim_{n\to\infty} \frac{\mathrm{E}[LC_n]}{n} = \gamma^*,$$

where $\gamma^* \in (0, 1]$.

Proof. If $\mu = \pi$, the sequence (X, Y) is stationary and therefore by superadditivity and Fekete's lemma or Kingman's subadditivity theorem (see [19]) implies that

$$\lim_{n\to\infty} \frac{\mathrm{E}[LC_n]}{n} = \sup_{k\geq 1} \frac{\mathrm{E}[LC_k]}{k} = \gamma^*$$

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

for some $\gamma^* \in (0, 1]$. When $\mu \neq \pi$, a coupling technique will prove the result. Let $\overline{Z}$ be a Markov chain with initial and stationary distribution π and having the same transition matrix P as the chain Z. Assume, further, that the emission probabilities are the same for Z and $\overline{Z}$, and denote by $(\overline{Z}, (\overline{X}, \overline{Y}))$ the corresponding HMM. Next, consider the coupling $(Z, \overline{Z})$ where the two chains stay together after the first time i for which $Z_i = \overline{Z}_i$, and let τ be the meeting time of Z and $\overline{Z}$. Next, and throughout, let $X^{(n)} := (X_1, \ldots, X_n)$, and similarly for $Y^{(n)}$, $\overline{X}^{(n)}$, and $\overline{Y}^{(n)}$. Since $LCS(X^{(n)};Y^{(n)}) - LCS(\overline{X}^{(n)};\overline{Y}^{(n)}) \leq n$, then for any $K > 0$,

$$\begin{aligned}
&|\mathrm{E}[LCS(X^{(n)};Y^{(n)}) - LCS(\overline{X}^{(n)};\overline{Y}^{(n)})]| \\
&= |\mathrm{E}[[LCS(X^{(n)};Y^{(n)}) - LCS(\overline{X}^{(n)};\overline{Y}^{(n)})]\mathbf{1}_{\tau>K}] \\
&\quad + \mathrm{E}[[LCS(X^{(n)};Y^{(n)}) - LCS(\overline{X}^{(n)};\overline{Y}^{(n)})]\mathbf{1}_{\tau\leq K}]| \\
&\leq n\mathrm{P}(\tau > K) + K + |\mathrm{E}[[LCS^K(X^{(n)};Y^{(n)}) - LCS^K(\overline{X}^{(n)};\overline{Y}^{(n)})]\mathbf{1}_{\tau\leq K}]| \\
&\leq n\mathrm{P}(\tau > K) + K, \qquad (2.1)
\end{aligned}$$

where $LCS^K(\cdot\,;\cdot)$ is now the length of the longest common subsequences restricted to the letters X_i and Y_i for $i > K$, noting also that when $\tau \leq K$, then $LCS^K(X^{(n)};Y^{(n)})$ and $LCS^K(\overline{X}^{(n)};\overline{Y}^{(n)})$ are identically distributed. If $K \in (mk, m(k+1)]$, for some $m \geq 0$, by an argument going back to Doeblin [7] (see also [20]),

$$\begin{aligned}
&\mathrm{P}(\tau > K) \\
&\leq \mathrm{P}(Z_k \neq \overline{Z_k}, Z_{2k} \neq \overline{Z_{2k}}, \ldots, Z_{mk} \neq \overline{Z_{mk}}) \\
&= \mathrm{P}(Z_k \neq \overline{Z_k})\mathrm{P}(Z_{2k} \neq \overline{Z_{2k}} \mid Z_k \neq \overline{Z_k}) \cdots \mathrm{P}(Z_{mk} \neq \overline{Z_{mk}} \mid Z_{(m-1)k} \neq \overline{Z_{(m-1)k}}) \\
&\leq (1-\epsilon)^{m-1} \\
&\leq c\alpha^K, \qquad (2.2)
\end{aligned}$$

where $\alpha = \sqrt[k]{1-\epsilon} \in (0, 1)$ and $c = 1/(1-\epsilon)^2$. Therefore, τ is finite with probability one. Choosing $K = \sqrt{n}$ yields $\mathrm{P}(\tau > K) + K/n \to 0$ and finally $\mathrm{E}[LC_n]/n \to \gamma^*$ as $n \to \infty$.

Clearly, $\mathrm{E}[LC_n] \leq n$; to see that $\gamma^* > 0$, note first that, by aperiodicity and irreducibility, $P^k \geq \epsilon$ for some fixed k and $\epsilon > 0$, i.e. all the entries of the matrix P^k are larger than some positive quantity ϵ. Therefore, $\mathrm{P}(X_1 = Y_{k+1}) > p$ for some $p = p(k, \epsilon) > 0$. Now,

$$LC_{nk+1} \geq \mathbf{1}_{X_1=Y_{k+1}} + \mathbf{1}_{X_{k+1}=Y_{2k+1}} + \cdots + \mathbf{1}_{X_{(n-1)k+1}=Y_{nk+1}}, \qquad (2.3)$$

and hence

$$\frac{np}{nk+1} \leq \frac{\mathrm{E}[LC_{nk+1}]}{nk+1}.$$

Letting $n \to \infty$ implies that $\gamma^* \in [p/(k+1), 1] \subset (0, 1]$, since $p > 0$. □

Remark 2.1. (i) Under a further assumption, we can show that $\gamma^* > \mathrm{P}(X_1 = Y_1)$. Indeed, assume that for all $x, y \in \mathcal{A}$, $z \in \mathcal{S}$, $\mathrm{P}(X_i = x, Y_i = y \mid Z_i = z) = \mathrm{P}(X_i = y, Y_i = x \mid Z_i = z) > 0$,

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

and let Z be started at the stationary distribution. Then, for any $n \geq 2$,

$$\begin{aligned}
\mathrm{E}[LC_n] &\geq \mathrm{E}[LC_{n-2}\mathbf{1}_{X_n=Y_n, X_{n-1}=Y_{n-1}}] + 2\mathrm{P}(X_n = Y_n, X_{n-1} = Y_{n-1}) \\
&\quad + \mathrm{E}[LC_{n-2}\mathbf{1}_{X_n=Y_n, X_{n-1}\neq Y_{n-1}}] + \mathrm{P}(X_n \neq Y_n, X_{n-1} = Y_{n-1}) \\
&\quad + \mathrm{E}[LC_{n-2}\mathbf{1}_{X_n\neq Y_n, X_{n-1}= Y_{n-1}}] + \mathrm{P}(X_n = Y_n, X_{n-1} \neq Y_{n-1}) \\
&\quad + \mathrm{E}[LC_{n-2}\mathbf{1}_{X_n\neq Y_n, X_{n-1}\neq Y_{n-1}}] + \mathrm{P}(X_n \neq Y_n, X_{n-1} \neq Y_{n-1}, X_n = Y_{n-1}) \\
&> \mathrm{E}[LC_{n-2}] + \mathrm{P}(X_n = Y_n) + \mathrm{P}(X_{n-1} = Y_{n-1}) \\
&= \mathrm{E}[LC_{n-2}] + 2\mathrm{P}(X_1 = Y_1)
\end{aligned}$$

by stationarity. Therefore, iterating, still using stationarity, and since $\mathrm{E}[LC_0] = 0$ while $\mathrm{E}[LC_1] = \mathrm{P}(X_1 = Y_1)$, it follows that for $n \geq 2$, $\mathrm{E}[LC_n] > n\mathrm{P}(X_1 = Y_1)$. Finally,

$$\gamma^* > \mathrm{P}(X_1 = Y_1) = \sum_{\alpha\in\mathcal{A}} \mathrm{P}(X_1 = \alpha)\mathrm{P}(Y_1 = \alpha),$$

and this inequality is strict since Fekete's lemma (see, e.g., [19]) ensures that $\gamma^* = \sup_n \mathrm{E}[LC_n]/n$.

(ii) Steele's general result, see [18], asserts that Proposition 2.1 holds if there is a stationary ergodic coupling for (X, Y). Such an example is when the sequences X and Y are generated by two independent aperiodic, homogeneous, and irreducible hidden Markov chains with the same parameters (and so the same emission probabilities). Indeed, at first, when the hidden chains Z_X and Z_Y generating respectively X and Y are started at the stationary distribution, convergence of $\mathrm{E}[LC_n]/n$ towards γ^* follows from superadditivity and Fekete's lemma (see [19]). As previously, $\gamma^* > 0$ since the properties of the hidden chains imply (2.3). Then, when the initial distribution is not the stationary distribution, we can proceed with arguments as above. In particular, let τ_1 and τ_2 be the respective meeting times of the chains $(Z_X, \overline{Z_X})$ and $(Z_Y, \overline{Z_Y})$, and let $\tau = \max(\tau_1, \tau_2)$. Then equation (2.1) continues to hold:

$$\begin{aligned}
|\mathrm{E}[LCS(X;Y) - LCS(\overline{X};\overline{Y})]| &\leq n\mathrm{P}(\tau > K) + K \\
&\leq 2n\mathrm{P}(\tau_1 > K) + K. \qquad (2.4)
\end{aligned}$$

Taking $K = \sqrt{n}$ and noting the exponential decay of $\mathrm{P}(\tau_1 > K)$ finishes the corresponding proof.

3. Rate of convergence

The previous section gives a mean convergence result; we now deal with its rate. Again, let (X, Y) be the outcome of a hidden Markov chain Z with μ, P, and π as initial distribution, transition matrix, and stationary distribution respectively. In this section we impose the additional restriction that the emission distributions for all states in the hidden chain are symmetric (this is discussed further in Proposition 3.1 and in the Appendix): for all $x, y \in \mathcal{A}$ and all $z \in \mathcal{S}$, $\mathrm{P}(X_i = x, Y_i = y \mid Z_i = z) = \mathrm{P}(X_i = y, Y_i = x \mid Z_i = z)$. Symmetry clearly implies that the conditional law of X given Z and of Y given Z are the same since, for all x, y and z,

$$\begin{aligned}
\mathrm{P}(X_i = x \mid Z_i = z) &= \sum_{y\in\mathcal{A}} \mathrm{P}(X_i = x, Y_i = y \mid Z_i = z) = \sum_{y\in\mathcal{A}} \mathrm{P}(X_i = y, Y_i = x \mid Z_i = z) \\
&= \mathrm{P}(Y_i = x \mid Z_i = z).
\end{aligned}$$

In turn this implies that X_i and Y_i are identically distributed.

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

Moreover, we need to control the dependency between X and Y, and a way to do so is via the β-mixing coefficient as given in Definition 3.3 of [5], which we now recall.

Definition 3.1. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two σ-fields $\subset \mathcal{F}$. Then the β-mixing coefficient associated with these sub-σ-fields of $\mathcal{F}$ is given by:

$$\beta(\mathcal{F}_1, \mathcal{F}_2) := \frac{1}{2} \sup \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathrm{P}(A_i \cap B_j) - \mathrm{P}(A_i)\mathrm{P}(B_j)|,$$

where the supremum is taken over all pairs of finite partitions $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_J\}$ of Ω such that $A_i \in \mathcal{F}_1$, for all $i \in \{1, \ldots, I\}$, $I \geq 1$ and $B_j \in \mathcal{F}_2$ for all $j \in \{1, \ldots, J\}$, $J \geq 1$.

In our case the above notion of the β-mixing coefficient is adopted for the σ-fields generated by *two* sequences. Moreover, by [5, Proposition 3.21], for a fixed $n \geq 1$, and since $X^{(n)} = (X_1, \ldots, X_n)$ and $Y^{(n)} = (Y_1, \ldots, Y_n)$ are discrete random vectors,

$$\begin{aligned}\beta(n) &:= \beta(\sigma(X^{(n)}), \sigma(Y^{(n)})) \\ &= \frac{1}{2} \sum_{u \in \mathcal{A}^n} \sum_{v \in \mathcal{A}^n} |\mathrm{P}(X^{(n)} = u, Y^{(n)} = v) - \mathrm{P}(X^{(n)} = u)\mathrm{P}(Y^{(n)} = v)|,\end{aligned}$$

where $\sigma(X^{(n)})$ and $\sigma(Y^{(n)})$ are the σ-fields generated by $X^{(n)}$ and $Y^{(n)}$. Clearly $X^{(n)}$ and $Y^{(n)}$ are independent if and only if $\beta(n) = 0$. Further, set $\beta^* := \lim_{n\to\infty} \beta(n)$, where the limit exists since $\beta(n)$ is non-decreasing in n and $\beta(n) \in [0, 1]$ (see [5, Section 5]).

Remark 3.1. (i) Another definition of the β-mixing coefficient based on 'past' and 'future' is often studied in the literature: see, for instance, [4, Section 2]. For a single sequence of random variables $S = (S_k)_{k\in\mathbb{Z}}$ and for $-\infty \leq J \leq L \leq \infty$, let

$$\mathcal{F}_J^L := \sigma(S_k, J \leq k \leq L),$$

and for each $n \geq 1$, let

$$\beta_n := \sup_{j\in\mathbb{Z}} \beta(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+n}^{\infty}).$$

In particular, [4, Theorem 3.2] implies that if S is a strictly stationary, finite-state Markov chain that is also irreducible and aperiodic, $\beta_n \to 0$ as $n \to \infty$. The mixing definition relevant to the approach here is different and this limiting behavior does not follow. A further discussion of the values of $\beta(n)$ is included in Remark 3.3 (i).

(ii) One might also be interested to use the α-mixing coefficient, defined for σ-fields $\mathcal{S}$ and $\mathcal{T}$ as:

$$\alpha(\mathcal{S}, \mathcal{T}) = 2 \sup\{|\mathrm{Cov}(\mathbf{1}_S, \mathbf{1}_T)| : (S, T) \in \mathcal{S} \times \mathcal{T}\}.$$

Suppose further that $\mathcal{T}$ has exactly N atoms. The following holds (see [4] and [3, Theorem 1]):

$$2\alpha(\mathcal{S}, \mathcal{T}) \leq \beta(\mathcal{S}, \mathcal{T}) \leq (8N)^{1/2}\alpha(\mathcal{S}, \mathcal{T}).$$

However, for this setting the number of atoms N will be $|\mathcal{A}|^n$, and since $\alpha(n) := \alpha(\sigma(X^{(n)}), \sigma(Y^{(n)}))$ is increasing, a bound on $\beta(n)$ using the inequality above is useless.

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

The following rate of convergence is the main result:

Theorem 3.1. *Let $(Z, (X, Y))$ be a hidden Markov model, where the sequence Z is an aperiodic time-homogeneous and irreducible Markov chain with finite state space $\mathcal{S}$. Let the distribution of the pairs (X_i, Y_i), $i = 1, 2, 3, \ldots$, be symmetric for all states in Z. Then, for all $n \geq 2$,*

$$\frac{\mathrm{E}[LC_n]}{n} \geq \gamma^* - 2\beta^* - C\sqrt{\frac{\ln n}{n}} - \frac{2}{n} - (1 - \mathbf{1}_{\mu=\pi})\Big(\frac{1}{\sqrt{n}} + c\alpha^{\sqrt{n}}\Big), \tag{3.1}$$

where $\alpha \in (0, 1)$, $c > 0$ are constants as in (2.2) and $C > 0$. All constants depend on the parameters of the model but not on n. Moreover, with the same α and c,

$$\frac{\mathrm{E}[LC_n]}{n} \leq \gamma^* + (1 - \mathbf{1}_{\mu=\pi})\Big(\frac{1}{\sqrt{n}} + c\alpha^{\sqrt{n}}\Big). \tag{3.2}$$

A key ingredient in proving Theorem 3.1 is a Hoeffding-type inequality for Markov chains, a particular case of a result due to Paulin [14], which is now recalled. It relies on the mixing time $\tau(\epsilon)$ of the Markov chain Z given by

$$\tau(\epsilon) := \min\{t \in \mathbb{N} : \bar{d}_Z(t) \leq \epsilon\},$$

where

$$\bar{d}_Z(t) := \max_{1 \leq i \leq N-t} \sup_{x,y \in \Lambda_i} d_{\mathrm{TV}}(\mathcal{L}(Z_{i+t} \mid Z_i = x), \mathcal{L}(Z_{i+t} \mid Z_i = y)),$$

and where $d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2}\sum_{x \in \Omega} |\mu(x) - \nu(x)|$ is the total variation distance between the two probability measures μ and ν on the finite set Ω.

Lemma 3.1. *Let $M := (M_1, \ldots, M_N)$ be a (not necessarily time-homogeneous) Markov chain, taking values in a Polish space $\Lambda = \Lambda_1 \times \cdots \times \Lambda_N$, with mixing time $\tau(\epsilon)$, $0 \leq \epsilon \leq 1$. Let*

$$\tau_{\min} := \inf_{0 \leq \epsilon < 1} \tau(\epsilon)\Big(\frac{2 - \epsilon}{1 - \epsilon}\Big)^2,$$

and let $f : \Lambda \to \mathbb{R}$ be such that there is $c \in \mathbb{R}_+^N$ with $|f(u) - f(v)| \leq \sum_{i=1}^N c_i \mathbf{1}_{u_i \neq v_i}$. Then, for any $t \geq 0$,

$$\mathrm{P}(f(M) - \mathrm{E}f(M) \geq t) \leq \exp\Big(\frac{-2t^2}{\tau_{\min}\sum_{i=1}^N c_i^2}\Big). \tag{3.3}$$

For our purposes, the Hoeffding-type inequality used below follows directly from (3.3) once we note that $(Z_i, X_i, Y_i)_{i \geq 1}$ is jointly a Markov chain on a bigger state space. Let $\tau(\epsilon)$ be the mixing time of this chain. Taking f to be the length of the longest common subsequences of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ we have $c = ((0, \ldots, 0), (1, \ldots, 1)) \in \mathbb{R}^n \times \mathbb{R}^{2n}$, since f is a function of Z, X, and Y whose values do not depend on Z. Letting $A := \sqrt{\tau_{\min}/2}$, (3.3) becomes

$$\mathrm{P}(LC_n - \mathrm{E}[LC_n] \geq t) \leq \exp\Big[\frac{-t^2}{A^2 n}\Big] \tag{3.4}$$

for all $t \geq 0$.

Remark 3.2. (i) When X and Y are generated by two independent hidden chains Z^X and Z^Y, the same reasoning yields (3.4), where now $\tilde{\tau}(\epsilon)$ is the mixing time of the chain $(Z_n^X, Z_n^Y, X_n, Y_n)_{n \geq 1}$.

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

(ii) The mixing time $\tau(\epsilon)$ of $(Z_n, X_n, Y_n)_{n\geq 1}$ is the same as the mixing time $\tilde{\tau}(\epsilon)$ of the chain $(Z_n)_{n\geq 1}$. Two proofs of this fact are provided in the Appendix.

Proof of Theorem 3.1. First, recall a result of Berbee [2] (see also [8, Theorem 1, Section 1.2.1], [16, Chapter 5], and [11]), asserting that on our probability space, which is rich enough, there exists $Y^{*(n)} := (Y_1^*, \ldots, Y_n^*)$, independent of $(Z, X)^{(n)} = ((Z_1, X_1), \ldots, (Z_n, X_n))$, having the same law as $Y^{(n)} = (Y_1, \ldots, Y_n)$ and such that

$$\mathrm{P}(Y^{(n)} \neq Y^{*(n)}) = \beta(n), \tag{3.5}$$

where $\beta(n) = \beta(\sigma((Z, X)^{(n)}), \sigma(Y^{(n)}))$ is the β-mixing coefficient of $(Z, X)^{(n)}$ and $Y^{(n)}$. Note also that if $(Y_i)_{i\geq 1}$ is stationary, then $(Y_1^*, \ldots, Y_k^*)$ and $(Y_\ell^*, \ldots, Y_{\ell+k-1}^*)$ are identically distributed for every $\ell, k \geq 1$, and that if $(X^{(n)}, Y^{(n)})$ is symmetric, then so is $(X^{(n)}, Y^{*(n)})$ where $X^{(n)} = (X_1, \ldots, X_n)$. Note finally that this implies that $Y^{*(n)}$ is independent of both $X^{(n)}$ and $Z^{(n)} = (Z_1, \ldots, Z_n)$.

Next, fix $k \in \mathbb{N}$; the idea of the proof is to relate $\mathrm{E}[LC_{kn}]$ to $\mathrm{E}[LC_{2n}]$. For $k = 4$, this is done in the i.i.d. case in [15]. However, we wish to take $k \to \infty$ and therefore follow arguments presented for the i.i.d. case in [12]. Call $(\nu, \tau) := (\nu_1, \ldots, \nu_r, \tau_1, \ldots \tau_r)$ an r-partition with $k \leq r \leq \lceil 2kn/(2n-1) \rceil$ if

$$\begin{aligned}
&1 = \nu_1 \leq \nu_2 \leq \cdots \leq \nu_{r+1} = kn + 1,\\
&1 = \tau_1 \leq \tau_2 \leq \cdots \leq \tau_{r+1} = kn + 1,\\
&(\nu_{j+1} - \nu_j) + (\tau_{j+1} - \tau_j) \in \{(2n-1, 2n\}, \ \text{for } j \in [1, r-1],\\
&(\nu_{r+1} - \nu_r) + (\tau_{r+1} - \tau_r) < 2n.
\end{aligned} \tag{3.6}$$

Let $\mathcal{B}_{k,n}^r$ be the set of all r-partitions defined as above, and let

$$\mathcal{B}_{k,n} = \bigcup_{r=k}^{\lceil 2kn/(2n-1) \rceil} \mathcal{B}_{k,n}^r.$$

If (ν, τ) is an r-partition, then setting

$$LC_{kn}(\nu, \tau) := \sum_{i=1}^{r} LCS(X_{\nu_i}, \ldots, X_{\nu_{i+1}-1}; Y_{\tau_i}, \ldots, Y_{\tau_{i+1}-1})$$

we have:

$$LC_{kn} = \max_{(\nu,\tau)\in\mathcal{B}(k,n)} LC_{kn}(\nu, \tau).$$

Let $\nu_{i+1} - \nu_i = n - m$, $\tau_{i+1} - \tau_i \leq n + m$ for $m \in (-n, n)$ and $\tau_i - \nu_i = \ell$. Then

$$\begin{aligned}
&\mathrm{E}[LCS(X_{\nu_i}, \ldots, X_{\nu_{i+1}-1}; Y_{\tau_i}, \ldots, Y_{\tau_{i+1}-1})]\\
&= \mathrm{E}[LCS(X_1, \ldots, X_{n-m}; Y_\ell, \ldots, Y_{\ell+n+m-1})] &(3.7)\\
&\leq \mathrm{E}\big[LCS(X_1, \ldots, X_{n-m}; Y_\ell^*, \ldots, Y_{\ell+n+m-1}^*)\mathbf{1}_{Y^{(kn)}=Y^{*(kn)}}\big]\\
&\quad + \min(n-m, n+m)\mathrm{P}\big(Y^{(kn)} \neq Y^{*(kn)}\big) &(3.8)\\
&\leq \mathrm{E}[LCS(X_1, \ldots, X_{n-m}; Y_\ell^*, \ldots, Y_{\ell+n+m-1}^*)] + n\beta(kn). &(3.9)
\end{aligned}$$

In the last expression the LCS is now a function of two independent sequences. Stationarity implies (3.7), and $LCS(X_1, \ldots, X_{n-m}; Y_\ell^*, \ldots, Y_{\ell+n+m-1}^*) \leq \min(n-m, n+m)$ leads to (3.8).

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

The error term $n\beta(kn)$ in (3.9) follows from an application of Berbee's result (3.5). The same properties also imply that

$$\begin{aligned}&\mathrm{E}[LCS(X_1,\ldots,X_{n-m};Y^*_\ell,\ldots,Y^*_{\ell+n+m-1})]\\&=\mathrm{E}[LCS(X_1,\ldots,X_{n-m};Y^*_1,\ldots,Y^*_{n+m})]\\&\le\mathrm{E}[LCS(X_1,\ldots,X_{n-m};Y_1,\ldots,Y_{n+m})]+n\beta(kn),\end{aligned}\tag{3.10}$$

and

$$\begin{aligned}&\mathrm{E}[LCS(X_1,\ldots,X_{n-m};Y^*_\ell,\ldots,Y^*_{\ell+n+m-1})]\\&=\mathrm{E}[LCS(X_1,\ldots,X_{n+m};Y^*_1,\ldots,Y^*_{n-m})]&(3.11)\\&\le\mathrm{E}[LCS(X_{n-m+1},\ldots,X_{2n};Y_{n+m+1},\ldots,Y_{2n})]+n\beta(kn),&(3.12)\end{aligned}$$

where the symmetry of the distributions of X and Y^* is used to get (3.11). Next, by the superadditivity of the LCSs as well as (3.9), (3.10), and (3.12),

$$\begin{aligned}&\mathrm{E}[LCS(X_{\nu_i},\ldots,X_{\nu_{i+1}-1};Y_{\tau_i},\ldots,Y_{\tau_{i+1}-1})]\\&\le\frac{1}{2}(\mathrm{E}[LCS(X_1,\ldots,X_{n-m};Y_1,\ldots,Y_{n+m})]\\&\quad+\mathrm{E}[LCS(X_{n-m+1},\ldots,X_{2n};Y_{n+m+1},\ldots,Y_{2n})]+2n\beta(kn))+n\beta(kn)\\&\le\frac{1}{2}(\mathrm{E}[LC_{2n}]+2n\beta(kn))+n\beta(kn)\\&=\frac{1}{2}\mathrm{E}[LC_{2n}]+2n\beta(kn).\end{aligned}\tag{3.13}$$

This inequality is key to the proof, since it yields an upper bound on $\mathrm{E}[LC_{kn}(\nu,\tau)]$ in terms of $\mathrm{E}[LC_{2n}]$, a quantity that does not depend on the partitioning (ν,τ). A similar result is central to the proof of the rate of convergence in the independent setting [1]. However, independence allows us to get (3.13) directly without the mere presence of or the need to introduce β-mixing coefficients. Moreover, the approach here is more direct. Applying Hoeffding's inequality and summing over all partitions provide a relation between $\mathrm{E}[LC_{kn}]$ and $\mathrm{E}[LC_{2n}]$ which can be used to get the rate of convergence. Indeed,

$$\mathrm{E}[LC_{kn}(\nu,\tau)]\le\frac{r}{2}(\mathrm{E}[LC_{2n}]+4n\beta(kn))\le\frac{1}{2}\left\lceil\frac{2kn}{2n-1}\right\rceil(\mathrm{E}[LC_{2n}]+4n\beta(kn)).$$

In addition, for $t>0$,

$$\begin{aligned}&\mathrm{P}\Big(LC_{kn}(\nu,\tau)-\frac{1}{2}\left\lceil\frac{2kn}{2n-1}\right\rceil(\mathrm{E}[LC_{2n}]+4n\beta(kn))>tkn\Big)\\&\le\mathrm{P}(LC_{kn}(\nu,\tau)-\mathrm{E}[LC_{kn}(\nu,\tau)]>tkn)\\&\le\exp\Big[-\frac{t^2kn}{A^2}\Big],\end{aligned}$$

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

where the second inequality follows from Lemma 3.1. Next, note that

$$\begin{aligned}
&\mathrm{P}\Big(LC_{kn} - \frac{1}{2}\Big\lceil \frac{2kn}{2n-1}\Big\rceil(\mathrm{E}[LC_{2n}] + 4n\beta(kn)) > tkn\Big) \\
&= \sum_{(\nu,\tau)\in\mathcal{B}_{k,n}} \mathrm{P}\Big(LC_{kn}(\nu,\tau) - \frac{1}{2}\Big\lceil \frac{2kn}{2n-1}\Big\rceil(\mathrm{E}[LC_{2n}] + 4n\beta(kn)) > tkn\Big) \\
&\le |\mathcal{B}_{k,n}| \exp\Big[-\frac{t^2kn}{A^2}\Big].
\end{aligned}$$

The above can be rewritten as:

$$\mathrm{P}\Big(\frac{LC_{kn}}{kn} > t + \frac{1}{k}\Big\lceil \frac{2kn}{2n-1}\Big\rceil\Big(\frac{\mathrm{E}[LC_{2n}]}{2n} + 2\beta(kn)\Big)\Big) \le |\mathcal{B}_{k,n}| \exp\Big[-\frac{t^2kn}{A^2}\Big].$$

Then, since $LC_{kn} \le kn$,

$$\begin{aligned}
\mathrm{E}\Big[\frac{LC_{kn}}{kn}\Big] &\le t + \frac{1}{k}\Big\lceil \frac{2kn}{2n-1}\Big\rceil\Big(\frac{\mathrm{E}[LC_{2n}]}{2n} + 2\beta(kn)\Big) \\
&\quad + \mathrm{P}\Big(\frac{LC_{kn}}{kn} > t + \frac{1}{k}\Big\lceil \frac{2kn}{2n-1}\Big\rceil\frac{\mathrm{E}[LC_{2n}]}{2n}\Big) \\
&\le t + \frac{1}{k}\Big\lceil \frac{2kn}{2n-1}\Big\rceil\Big(\frac{\mathrm{E}[LC_{2n}]}{2n} + 2\beta(kn)\Big) + |\mathcal{B}_{k,n}| \exp\Big[-\frac{t^2kn}{A^2}\Big]. \qquad (3.14)
\end{aligned}$$

Next, a bound on $|\mathcal{B}_{k,n}|$ is obtained using methods as in [12]. Recall that $k \le r \le \lceil 2kn/(2n-1)\rceil$ and that $\mathcal{B}_{k,n} = \bigcup_{r=k}^{2kn/2n-1} \mathcal{B}_{k,n}^r$. Now

$$|\mathcal{B}_{k,n}^r| \le 2^{r-1}2n\binom{nk+r-1}{r-1}. \qquad (3.15)$$

Indeed, the sizes of the partitions on the X side should sum to nk, which gives a factor of less than $\binom{nk+r-1}{r-1}$. Also, for each choice of the first $r-1$ elements of the partition on the X side we have at most 2 choices on the Y side. The last interval can take at most $2n$ values, as per (3.6). Recall Stirling's formula (see [10]): for $n \ge 1$,

$$n^n\mathrm{e}^{-n}\sqrt{2\pi n}\mathrm{e}^{1/(12n+1)} \le n! \le n^n\mathrm{e}^{-n}\sqrt{2\pi n}\mathrm{e}^{1/12n}.$$

Since in the end of the proof $k \to \infty$, this bound can be used in (3.15) to obtain:

$$\begin{aligned}
|\mathcal{B}_{k,n}^r| &\le (2^{r-1}2n)\frac{(nk+r-1)^{nk+r-1}\sqrt{2\pi(nk+r-1)}\mathrm{e}^{1/12(nk+r-1)}}{(r-1)(r-1)\sqrt{2\pi(r-1)}\mathrm{e}^{1/12(r-1)+1}(nk)^{nk}\sqrt{2\pi nk}\mathrm{e}^{1/12(nk)+1}} \\
&\le 2^r n\frac{(nk+r-1)^{nk+r-1}}{(r-1)^{r-1}(nk)^{nk}} \\
&\le 2^r n\Big(1+\frac{nk}{r-1}\Big)^{r-1}\Big(1+\frac{2}{2n-1}\Big)^{nk} \\
&\le 2^r n\Big(1+n+\frac{n}{k-1}\Big)^{\frac{2nk}{2n-1}}\Big(\frac{2n+1}{2n-1}\Big)^{nk}.
\end{aligned}$$

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

The last inequality in the above expression holds true since $k \le r \le \lceil 2kn/(2n-1) \rceil$. Then, for $|\mathcal{B}_{k,n}|$ we get:

$$\begin{aligned}
|\mathcal{B}_{k,n}| &\le \Big(\frac{2nk}{2n-1} - k + 2\Big) \max_r |\mathcal{B}^r_{k,n}| \\
&\le \Big(\frac{k}{2n-1} + 2\Big) 2^r n \Big(1 + n + \frac{n}{k-1}\Big)^{\frac{2nk}{2n-1}} \Big(\frac{2n+1}{2n-1}\Big)^{nk} \\
&\le \exp\Big(\Big(\frac{\ln\left(\frac{k}{2n-1} + 2\right)}{nk} + \frac{r\ln 2 + \ln n}{nk} + \frac{2}{2n-1}\ln(2n) + \ln\Big(\frac{2n+1}{2n-1}\Big)\Big)nk\Big) \\
&\le \exp\Big(\Big(\frac{\ln k}{k} + \frac{2}{2n-1}\ln 2 + \frac{2}{2n-1}\ln(2n) + \ln\Big(\frac{2n+1}{2n-1}\Big)\Big)nk\Big) \\
&\le \exp\Big(\Big(\frac{\ln k}{k} + \frac{4}{2n-1}\ln 2 + \frac{2}{2n-1}\ln n + \ln\Big(\frac{2n+1}{2n-1}\Big)\Big)nk\Big) \\
&\le \exp(10k\ln n),
\end{aligned}$$

where the last inequality holds for large k, in particular $k > n$, and since $\ln(1+x) \le x$ for $x > 0$. Let $t = 2A\sqrt{10}\sqrt{\ln n/n}$. Then

$$\begin{aligned}
|\mathcal{B}_{k,n}| \exp\Big(-\frac{t^2 kn}{A^2}\Big) &\le \exp(10k\ln n)\exp\Big(-\frac{t^2 kn}{A^2}\Big) \\
&\le \exp(-30k\ln n).
\end{aligned}$$

Next, note that, as $k \to \infty$, $\mathrm{E}[LC_{kn}/(kn)] \to \gamma^*$ and that

$$\frac{1}{k}\Big\lceil \frac{2kn}{2n-1} \Big\rceil \le \frac{1}{k}\Big(\frac{2kn}{2n-1} + 1\Big) \to \frac{2n}{2n-1}.$$

Recall also that $\beta^* = \lim_{n\to\infty}\beta(n) = \lim_{k\to\infty}\beta(kn)$. Then (3.14) implies that

$$\frac{2n}{2n-1}\Big(\frac{\mathrm{E}[LC_{2n}]}{2n} + 2\beta^*\Big) \ge \gamma^* - 2A\sqrt{10}\sqrt{\frac{\ln n}{n}}, \tag{3.16}$$

and, finally,

$$\begin{aligned}
\frac{\mathrm{E}[LC_{2n}]}{2n} &\ge \frac{2n-1}{2n}\Big(\gamma^* - 2A\sqrt{10}\sqrt{\frac{\ln n}{n}}\Big) - 2\beta^* \\
&\ge \gamma^* - 2\beta^* - 2A\sqrt{10}\sqrt{\frac{\ln n}{n}} - \frac{1}{2n}.
\end{aligned}$$

To get the result for words of odd length note that, by (3.16),

$$\begin{aligned}
\frac{\mathrm{E}[LC_{2n+1}]}{2n+1} &\ge \frac{\mathrm{E}[LC_{2n}]}{2n+1} \\
&\ge \frac{2n-1}{2n+1}\Big(\gamma^* - 2A\sqrt{10}\sqrt{\frac{\ln n}{n}}\Big) - \frac{2n}{2n+1}2\beta^* \\
&\ge \gamma^* - 2\beta^* - 2A\sqrt{10}\sqrt{\frac{\ln n}{n}} - \frac{2}{2n+1}.
\end{aligned}$$

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

Of course, these last bounds are only of interest, for n large enough, if $\gamma^* > 2\beta^*$. Otherwise we get the trivial lower bound 0 (see Remark 3.3 below). We are then left with slightly modifying the constants to get (3.1). The extra term on the right-hand side in (3.1) accounts for the difference in initial distributions (2.1).

The proof of the upper bound (3.2), where symmetry is not needed, follows by combining Fekete's lemma (see [19]) with (2.1) and (2.2). □

Remark 3.3. (i) Recall that the β-mixing coefficient $\beta(n)$ is a measure on the dependency between $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$. The bounds in Theorem 3.1 rely on $\beta^* := \lim_{n\to\infty} \beta(n)$, which somehow quantifies a weak dependency requirement, and $\beta^* \neq 0$ unless the sequences X and Y are independent. Note also that the lower bound in Theorem 3.1 is meaningful only if $2\beta^* < \gamma^*$. Besides the independent case, there are instances for which this condition is satisfied. For example, let X and Y both be Markov chains with L states and with the same transition matrix P, where some rows of P are equal to $(1, 1, 1, \dots, 1)/L$, i.e. such that there exists a set of states $\mathcal{L}$ such that the transition probability between each one of these states is uniform. Let the initial distribution of X_1 be μ with $\mu(x) = 0$ if $x \notin \mathcal{L}$ and assume that $Y_1 = X_1$. Then the sequence $\tilde{Y}$ defined, for all n, via $\tilde{Y}_i = Y_i$ for $i \geq 1$, while Y_1 is distributed according to μ, will be such that $\tilde{Y}^{(n)}$ and $Y^{(n)}$ have the same distribution. Moreover, for all n, $\tilde{Y}^{(n)}$ and $X^{(n)}$ will be independent and $\mathrm{P}(\tilde{Y}^{(n)} \neq Y^{(n)}) \geq \beta(n)$, but $\mathrm{P}(\tilde{Y}^{(n)} \neq Y^{(n)}) = \mathrm{P}(Y_1 \neq \tilde{Y}_1)$, which can be made as small as desired for a suitable choice of μ. Thus the lower bound in Theorem 3.1 holds and is meaningful.

(ii) There are instances when the lower bound in Theorem 3.1 is vacuous. Such a case is when $X_i = Y_i$ for all $i \geq 1$ and the X_i are independent and uniformly distributed over the letters in $\mathcal{A}$. Then it is clear that $\gamma^* = 1$, whereas we can show that

$$\beta(n) = 1 - \frac{1}{|\mathcal{A}|^n},$$

and so $\beta^* = 1$. In this case the lower bound in (3.1) is a negative quantity.

(iii) Theorem 3.1 continues to hold for Markov chains with a general state space Λ. Indeed, the Hoeffding inequality (3.4) is true when Λ is a Polish space. The exponential decay (2.2) holds when Λ is *petite*, i.e. when there exist a positive integer n_0, $\epsilon > 0$, and a probability measure ν on Λ such that $P^{n_0}(x, A) \geq \epsilon\nu(A)$ for every measurable A and $x \in \Lambda$, and where $P^{n_0}(x, A)$ is the n_0-step transition law of the Markov chain (see [17, Theorem 8]).

When X and Y are generated by independent hidden Markov models the following variant of Theorem 3.1 holds (for a sketch of the proof, see the Appendix).

Corollary 3.1. *Let (Z_X, X) and (Z_Y, Y) be two independent hidden Markov models, where the latent chains Z_X and Z_Y have the same initial distribution, transition matrix, and emission probabilities. Then, for all $n \geq 2$,*

$$\frac{\mathrm{E}[LC_n]}{n} \geq \gamma^* - C\sqrt{\frac{\ln n}{n}} - \frac{2}{n} - (1 - \mathbf{1}_{\mu=\pi})\Big(\frac{1}{\sqrt{n}} + c\alpha^{\sqrt{n}}\Big),$$

where $\alpha \in (0, 1)$, $c > 0$ are constants as in (2.2) and $C > 0$. All constants depend on the parameters of the model but not on n. Moreover, with the same α and c,

$$\frac{\mathrm{E}[LC_n]}{n} \leq \gamma^* + (1 - \mathbf{1}_{\mu=\pi})\Big(\frac{1}{\sqrt{n}} + c\alpha^{\sqrt{n}}\Big).$$

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

As mentioned at the end of the proof of Theorem 3.1, the symmetry of the distribution of (X_i, Y_i) is used only for proving the lower bound. Let

$$h(n) := \max_{m\in[-n,n]} \left(2\sum_{i=1}^{n-m} \mathrm{P}(X_i \neq Y_i) + \sum_{i=n-m+1}^{n+m} \mathrm{P}(X_i \neq Y_i) \right).$$

Then the following result holds.

Proposition 3.1. *Let $(Z, (X, Y))$ be a hidden Markov model, where the sequence Z is an aperiodic time-homogeneous and irreducible Markov chain with finite state space $\mathcal{S}$. Then, for all $n \geq 2$,*

$$\frac{\mathrm{E}[LC_n]}{n} \geq \gamma^* - \frac{h(n)}{n} - 2\beta^* - C\sqrt{\frac{\ln n}{n}} - \frac{2}{n} - (1 - \mathbf{1}_{\mu=\pi})\left(\frac{1}{\sqrt{n}} + c\alpha^{\sqrt{n}}\right).$$

For a sketch of proof of this proposition, and some comments on $h(n)$, we again refer the reader to the Appendix.

Appendix A

First, as asserted in Remark 3.2 (ii), this appendix provides two proofs of the fact that the mixing time $\tau(\epsilon)$ of $(Z_n, X_n, Y_n)_{n\geq 1}$ is the same as the mixing time $\tilde{\tau}(\epsilon)$ of the chain $(Z_n)_{n\geq 1}$.

Proof 1. let $\tilde{T} = (\tilde{T}_n)_{n\geq 1}$ be a Markov chain with finite state space $\mathcal{S}$. Each $\tilde{T}_i$ emits an observed variable T_i according to some probability distribution that depends only on the state $\tilde{T}_i$. Let $T = (T_n)_{n\geq 1}$ and assume $T_i \in \mathcal{A}$, a finite alphabet. Note that $(\tilde{T}, T)$ is a Markov chain; let $\tau(\epsilon)$ be its mixing time, and let $\tilde{\tau}(\epsilon)$ be the mixing time for the hidden chain $\tilde{T}$. Then

$$\begin{aligned}
&d_{\mathrm{TV}}(\mathcal{L}((\tilde{T}_{i+t}, T_{i+t}) \mid (\tilde{T}_i, T_i) = (x, u)), \mathcal{L}((\tilde{T}_{i+t}, T_{i+t}) \mid (\tilde{T}_i, T_i) = (y, v)))\\
&= \frac{1}{2} \sum_{(z,w)\in\mathcal{S}\times\mathcal{A}} |\mathrm{P}((\tilde{T}_{i+t}, T_{i+t}) = (z, w) \mid (\tilde{T}_i, T_i) = (x, u)) -\\
&\quad - \mathrm{P}((\tilde{T}_{i+t}, T_{i+t}) = (z, w) \mid (\tilde{T}_i, T_i) = (y, v))|\\
&= \frac{1}{2} \sum_{(z,w)} |\mathrm{P}(\tilde{T}_{i+t} = z \mid \tilde{T}_i = x)\mathrm{P}(z \to w) - \mathrm{P}(\tilde{T}_{i+t} = z \mid \tilde{T}_i = y)\mathrm{P}(z \to w)|\\
&= \frac{1}{2} \sum_{(z,w)} \mathrm{P}(z \to w)|\mathrm{P}(\tilde{T}_{i+t} = z \mid \tilde{T}_i = x) - \mathrm{P}(\tilde{T}_{i+t} = z \mid \tilde{T}_i = y)|\\
&= \frac{1}{2} \sum_{z\in\mathcal{S}} |\mathrm{P}(\tilde{T}_{i+t} = z \mid \tilde{T}_i = x) - \mathrm{P}(\tilde{T}_{i+t} = z \mid \tilde{T}_i = y)|\\
&= d_{\mathrm{TV}}(\mathcal{L}(\tilde{T}_{i+t} \mid \tilde{T}_i = x), \mathcal{L}(\tilde{T}_{i+t} \mid \tilde{T}_i = y)),
\end{aligned}$$

where $\mathrm{P}(z \to w) := \mathrm{P}(T_i = w | \tilde{T}_i = z)$, i.e. the probability that a state with value $z \in \mathcal{S}$ emits $w \in \mathcal{A}$. By the definitions of T and $\tilde{T}$ this last probability does not depend on i. Then $\sum_{w\in\mathcal{A}} \mathrm{P}(z \to w) = 1$. Therefore, $\overline{d}_{(\tilde{T},T)}(t) = \overline{d}_{\tilde{T}}(t)$ and $\tau(\epsilon) = \tilde{\tau}(\epsilon)$. □

Proof 2. An alternative approach to proving the result of Remark 3.2 (ii) relies on coupling arguments and was kindly suggested by D. Paulin in personal communications with the authors. First, recall the following classical result [13, Proposition 4.7].

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

Lemma A.1. *Let μ and ν be two probability distributions on Ω. Then*

$$d_{\mathrm{TV}}(\mu, \nu) = \inf\{\mathrm{P}(X \neq Y) : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}.$$

Moreover, there is a coupling (X, Y) which attains the infimum and such a coupling is called optimal.

Let $(\tilde{T}^1, \tilde{T}^2)$ be an optimal coupling according to $d_{\mathrm{TV}}(\mathcal{L}(\tilde{T}_t \mid \tilde{T}_1 = x), \mathcal{L}(\tilde{T}_t \mid \tilde{T}_1 = y))$ for some $x, y \in \mathcal{S}$, i.e. $\tilde{T}^1$ and $\tilde{T}^2$ are Markov chains with the same transition probability as $\tilde{T}$, $\tilde{T}^1_0 = x$, $\tilde{T}^2_0 = y$, and

$$\mathrm{P}(\tilde{T}^1_t \neq \tilde{T}^2_t) = \quad d_{\mathrm{TV}}(\mathcal{L}(\tilde{T}_t \mid \tilde{T}_1 = x), \mathcal{L}(\tilde{T}_t \mid \tilde{T}_1 = y)). \tag{A.1}$$

Next, let T^1_t and T^2_t be respectively distributed according to the distributions associated with $\tilde{T}^1_t$ and $\tilde{T}^2_t$ and be independent of all the other random variables. In addition, if for some $t \geq 1$, $\tilde{T}^1_t = \tilde{T}^2_t$, then $T^1_t = T^2_t$. Then

$$\mathrm{P}(\tilde{T}^1_t \neq \tilde{T}^2_t) = \mathrm{P}((\tilde{T}^1_t, T^1_t) \neq (\tilde{T}^2_t, T^2_t)),$$

and by Lemma A.1, for any $u, v \in \mathcal{A}$ and any $i \geq 1$,

$$\begin{aligned} &\mathrm{P}((\tilde{T}^1_t, T^1_t) \neq (\tilde{T}^2_t, T^2_t)) \\ \geq \quad & d_{\mathrm{TV}}(\mathcal{L}((\tilde{T}_{i+t}, T_{i+t}) \mid (\tilde{T}_i, T_i) = (x, u)), \mathcal{L}((\tilde{T}_{i+t}, T_{i+t}) \mid (\tilde{T}_i, T_i) = (y, v))). \end{aligned}$$

Together with (A.1), the above yields

$$\begin{aligned} & d_{\mathrm{TV}}(\mathcal{L}(\tilde{T}_t \mid \tilde{T}_1 = x), \mathcal{L}(\tilde{T}_t \mid \tilde{T}_1 = y)) \geq \\ \geq \quad & d_{\mathrm{TV}}(\mathcal{L}((\tilde{T}_{i+t}, T_{i+t}) \mid (\tilde{T}_i, T_i) = (x, u)), \mathcal{L}((\tilde{T}_{i+t}, T_{i+t}) \mid (\tilde{T}_i, T_i) = (y, v))). \end{aligned}$$

Taking the sup over x, y, u, v gives $\overline{d}_{(\tilde{T},T)}(t) \leq \overline{d}_{\tilde{T}}(t)$.

For the reverse inequality, consider the optimal coupling $((\tilde{T}^1, T^1), (\tilde{T}^2, T^2))$ according to $d_{\mathrm{TV}}(\mathcal{L}((\tilde{T}_t, T_t) \mid (\tilde{T}_1, T_1) = (x, u)), \mathcal{L}((\tilde{T}_t, T_t) \mid (\tilde{T}_1, T_1) = (y, v)))$, for some $x, y \in \mathcal{S}$ and $u, v \in \mathcal{A}$. Then

$$\mathrm{P}((\tilde{T}^1_t, T^1_t) \neq (\tilde{T}^2_t, T^2_t)) = d_{\mathrm{TV}}(\mathcal{L}((\tilde{T}_t, T_t) \mid (\tilde{T}_1, T_1) = (x, u)), \mathcal{L}((\tilde{T}_t, T_t) \mid (\tilde{T}_1, T_1) = (y, v))) \tag{A.2}$$

and

$$\mathrm{P}((\tilde{T}^1_t, T^1_t) \neq (\tilde{T}^2_t, T^2_t)) \geq \mathrm{P}(\tilde{T}^1_t \neq \tilde{T}^2_t).$$

However, by Lemma A.1, for any $i \geq 1$,

$$\mathrm{P}(\tilde{T}^1_t \neq \tilde{T}^2_t) \geq d_{\mathrm{TV}}(\mathcal{L}(\tilde{T}_{i+t} \mid \tilde{T}_i = x), \mathcal{L}(\tilde{T}_{i+t} \mid \tilde{T}_i = y)). \tag{A.3}$$

Taking the sup in (A.2), (A.3) gives $\overline{d}_{(\tilde{T},T)}(t) \geq \overline{d}_{\tilde{T}}(t)$, and then $\overline{d}_{(\tilde{T},T)}(t) = \overline{d}_{\tilde{T}}(t)$. □

Proof of Corollary 3.1. The Hoeffding inequality (3.4) holds as long as (Z, X, Y) is a Markov chain. In addition, (X, Y) has to be symmetric (see proof of Proposition 3.1) in order for (3.13) to hold. Again, one such setting is when X and Y are two independent HMMs

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

with the same transition matrix for the latent chain and same emission probabilities. A rate of convergence result then follows from arguments as in Section 3. The bound on $\mathcal{B}_{k,n}$ is the same, and there is a Hoeffding-type inequality for this model as per Remark 3.2 (i). One thing that differs is the bound in (3.13). In the present case it is much easier to get. When started at the stationary distribution, by stationarity, independence, and symmetry, we have:

$$\begin{aligned} LCS(X_{\nu_i}, \dots, X_{\nu_{i+1}-1}; Y_{\tau_i}, \dots, Y_{\tau_{i+1}-1}) &\le LCS(X_1, \dots, X_{n-m}, Y_1, \dots, Y_{n+m}) \\ &= LCS(X_1, \dots, X_{n+m}; Y_1, \dots, Y_{n-m}) \\ &\le \frac{1}{2} LC_{2n}. \end{aligned}$$

In particular, there is no need to introduce mixing coefficients in this case ($\beta = 0$). When the hidden chains are not started at the stationary distribution we get an error as in (2.4). Then, Theorem 3.1 holds but with constants depending on the new model. Moreover, this setting reduces to the one where X and Y are independent Markov chains by letting each state of the hidden chains emit a unique letter, which can further recover the i.i.d. case originally obtained in [1]. □

Proof of Proposition 3.1. The symmetry of the distribution of (X, Y) is only used to get (3.11), which implies that for any $m \in \{-n+1, \dots, n-1\}$, $LCS(X_1, \dots, X_{n-m}; Y_1, \dots, Y_{n+m})$ and $LCS(X_1, \dots, X_{n+m}; Y_1, \dots, Y_{n-m})$ are identically distributed and bounded above by half of LC_{2n}. Such a result yields a comparison between $\mathrm{E}[LC_{2n}]$ and $\mathrm{E}[LC_{kn}]$, leading as $k \to \infty$ to a lower bound on $\mathrm{E}[LC_{2n}]$ involving γ^*. Without assuming symmetry, the step (3.11) in obtaining (3.13) needs to be modified. One way to do so is to make use of the Lipschitz property of the LCS to get the following estimate:

$$\begin{aligned} &LCS(X_1, \dots, X_{n-m}; Y_1, \dots, Y_{n+m}) \\ &\quad = LCS(X_1, \dots, X_{n+m}; Y_1, \dots, Y_{n-m}) + (LCS(X_1, \dots, X_{n-m}; Y_1, \dots, Y_{n+m}) \\ &\qquad - LCS(X_1, \dots, X_{n+m}; Y_1, \dots, Y_{n-m})) \\ &\quad \le LCS(Y_1, \dots, Y_{n-m}; X_1, \dots, X_{n+m}) + 2 \sum_{i=1}^{n-m} \mathbf{1}_{X_i \ne Y_i} + \sum_{i=n-m+1}^{n+m} \mathbf{1}_{X_i \ne Y_i}. \end{aligned}$$

Taking expectations, (3.13) becomes

$$\mathrm{E}[LCS(X_{\nu_i}, \dots, X_{\nu_{i+1}-1}; Y_{\tau_i}, \dots, Y_{\tau_{i+1}-1})] \le \frac{1}{2}(\mathrm{E}[LC_{2n}] + h(n)) + 2n\beta(kn),$$

where

$$h(n) := \max_{m \in [-n,n]} \left(2 \sum_{i=1}^{n-m} \mathrm{P}(X_i \ne Y_i) + \sum_{i=n-m+1}^{n+m} \mathrm{P}(X_i \ne Y_i) \right).$$

This leads to a non-symmetric version of (3.1), namely

$$\frac{\mathrm{E}[LC_n]}{n} \ge \gamma^* - C\sqrt{\frac{\ln n}{n}} - \frac{h(n)}{n} - 2\beta^* - \frac{1}{n-2} - (1 - \mathbf{1}_{\mu=\pi})\Big(\frac{1}{\sqrt{n}} + c\alpha^{\sqrt{n}}\Big). \tag{A.4}$$

This completes the proof. □

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

If $h(n) = O(\sqrt{n \ln n})$ then the rate in (3.17) or (A.4) will be the same as in (3.1). Such will be the case when (Z', X) and (Z'', Y) are two independent hidden Markov models and $Z = (Z', Z'')$ is a coupling of the two latent chains such that if $Z'_i = Z''_i$ then $Z'_j = Z''_j$ for any $j > i$. Then, $(Z, (X, Y))$ is a hidden Markov model where $X_i = Y_i$ once the two latent chains have met, and, by (2.2), $h(n) = O(\sqrt{n \log n})$. However, $h(n)$ can be much larger, e.g. of order n. A case in point is when the X_i and Y_i are i.i.d. Bernoulli random variables with parameters $1/3$ and $1/2$ respectively. Then $\mathrm{P}(X_i \neq Y_i) = \mathrm{P}(X_i = 0, Y_i = 1) + \mathrm{P}(X_i = 1, Y_i = 0) = 1/6 + 2/6 = 1/2$, for all $i \geq 1$, and

$$\left(2\sum_{i=1}^{n-m} \mathrm{P}(X_i \neq Y_i) + \sum_{i=n-m+1}^{n+m} \mathrm{P}(X_i \neq Y_i)\right) = \big(2(n-m)1/2 + (2m)1/2\big) = n.$$

Note also that when $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ are independent sequences of random variables, the symmetry assumption is equivalent to X and Y being identically distributed.

Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant No. 1440140, while the first author was in residence at the Mathematical Sciences Research Institute in Berkeley, California, during the Fall semester of 2017. His research was also supported in part by grants #246283 and #524678 from the Simons Foundation. The second author was partially supported by the TRIAD NSF grant (award 1740776). Both authors are grateful to J. Spouge and S. Eddy for encouragement and correspondence on the relevance of the hidden Markov models in computational biology.

References

[1] ALEXANDER, K. (1994). The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Prob.* **4,** 1074–1082.
[2] BERBEE, H. C. P. (1979). *Random Walks with Stationary Increments and Renewal Theory*. Mathematical Centre Tracts, 112. Mathematisch Centrum, Amsterdam.
[3] BRADLEY, R. (1983) Approximation theorems for strongly mixing random variables. *Michigan Math. J.* **30,** 69–81.
[4] BRADLEY, R. (2005) Basic properties of strong mixing conditions. A survey and some open questions. *Prob. Surveys* **2,** 104–144.
[5] BRADLEY, R. (2007). *Introduction to Strong Mixing Conditions*, Vol. 1. Kendrick Press, Heber City, UT.
[6] CHVÁTAL, V. AND SANKOFF, D. (1975). Longest common subsequences of two random sequences. *J. Appl. Prob.* **12,** 306–315.
[7] DOEBLIN, W. (1938). Exposé de la théorie des chaînes simples constantes de Markoff à un nombre fini d'états, *Revue Math. de l'Union Interbalkanique* **2,** 77–105.
[8] DOUKHAN, P. (1994). *Mixing. Properties and Examples*. Lecture Notes Statist., 85. Springer, New York.
[9] DURBIN, R., EDDY, S., KROGH, A. AND MITCHISON, G. (1998). *Biological Sequence Analysis*. Cambridge University Press.
[10] FELLER, W. (1968). *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd edn. John Wiley, New York.
[11] GOLDSTEIN, S. (1979). Maximal coupling. *Z. Wahrsch. verw. Gebiete* **46,** 193–204.
[12] LEMBER, J., MATZINGER, H. AND TORRES, F. (2012). The rate of convergence of the mean score in random sequence comparison. *Ann. Appl. Prob.* **22,** 1046–1058.
[13] LEVIN, D., PERES, Y. AND WILMER, E. (2008). *Markov Chains and Mixing Times*. AMS, Providence, RI.
[14] PAULIN, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Prob.* **20,** no. 79.
[15] RHEE, W. (1995). On rates of convergence for common subsequences and first passage time. *Ann. Appl. Prob.* **5,** 44–48.

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press

[16] RIO, E. (2017) *Asymptotic Theory of Weakly Dependent Random Processes*. Probability Theory and Stochastic Modelling, 80. Springer, Berlin.
[17] ROBERTS, G. AND ROSENTHAL, J. (2004) General state space Markov chains and MCMC algorithms. *Prob. Surveys* **1,** 20–71.
[18] STEELE, M. (1982). Long common subsequences and the proximity of two random strings. *SIAM J. Appl. Math.* **42,** 731–737.
[19] STEELE, M.(1997). *Probability Theory and Combinatorial Optimization*. SIAM, Philadelphia, PA, 18–21.
[20] THORISSON, H. (2000). *Coupling, Stationarity and Regeneration*. Probability and its Applications. Springer, New York.

https://doi.org/10.1017/jpr.2019.26 Published online by Cambridge University Press