Low Complexity Sequential Lossless Coding for Piecewise Stationary Memoryless Sources *

Gil I. Shamir† and Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL

Abstract

Three strongly sequential, lossless compression schemes, one with linearly growing per-letter computational complexity, and two with fixed per-letter complexity, are presented and analyzed for memoryless sources with abruptly changing statistics. The first method, which improves on Willems' weighting approach, asymptotically achieves a lower bound on the redundancy, and hence is optimal. The second scheme achieves redundancy of $O(\log N/N)$ when the transitions in the statistics are large, and $O(\log\log N/\log N)$ otherwise. The third approach always achieves redundancy of $O\left(\sqrt{\log N/N}\right)$. Obviously, the two fixed complexity approaches can be easily combined to achieve the better redundancy between the two. Simulation results support the analytical bounds derived for all the coding schemes.

Index Terms: Piecewise stationary memoryless source, universal coding, redundancy, minimum description length, ideal code length, sequential coding, strongly sequential coding, transition path, weighting, segmentation, change detection, source block code.

*This work was supported by the Israeli Consortium of Ground Stations for Satellite Communications administered by the S. Neaman Institute for Science & Technology. This work partially summarizes the M.Sc. research work of the first author.

†Gil Shamir is presently at the Department of Electrical Engineering, 275 Fitzpatrick Hall, University of Notre Dame, Notre Dame, IN 46556, U.S.A.

1 Introduction

Traditional sequential universal lossless source coding schemes are usually designed for classes of stationary sources. Not surprisingly, these schemes may perform poorly when the source is non-stationary, unless some adaptation mechanism is applied. While adaptive schemes such as the dynamic Huffman code [4], [9], [15], [22] and variations of the sliding-window Lempel-Ziv algorithm [21], [24], [25] have been developed and applied for general non-stationary sources, much less attention has been devoted to systematic, rigorous theoretical development of universal codes for *simple* classes of non-stationary sources. One example of such a class is that of memoryless sources with piecewise fixed letter probabilities [11]-[12], [16]-[19], namely, sources for which the probability mass function (PMF) is subjected to occasional abrupt changes. This model is useful in several application areas, like compression of speech or text retrieved from several sources, edge information in images, and abrupt scene changes in video coding.

In this paper, we adopt this simple model of piecewise stationary memoryless source (*PSMS*). Neither the source parameters at any stationary segment, nor the transition locations and their number are assumed to be known in advance. One can show that traditional adaptation mechanisms combined with classical compression schemes perform poorly for this class. Dynamic Huffman coding requires large block length, and thus exponentially large dictionary, in order to approach the entropy even in a stationary segment. Variations of the Lempel-Ziv algorithm require increasing window length, which results in slow convergence to the source entropy. One may use adaptive entropy coding w.r.t. estimated letter probabilities across a sliding window or an exponential one. But such estimates have non-decaying variance, and thus yield poor coding performance.

This calls for a different approach. First, recall the well-known fact that, ignoring asymptotically negligible integer length constraints, the problem of sequential lossless coding (using e.g., arithmetic coding [7], [8]), is completely equivalent to the problem of sequential probability assignment, where the length function of the code is understood as the negative logarithm of the assigned probability, i.e., the *ideal code length*. From this point on, we treat the latter problem.

To the best of our knowledge, universal coding for the PSMS model was first investigated by Merhav [11] (see also [12]). Merhav showed that the average universal coding redundancy over all sequences of N letters, drawn by almost any PSMS with C transitions from an alphabet of r letters, is lower bounded by

$$R_N \geq (1-\varepsilon)\left(\frac{r-1}{2}(C+1)+C\right)\frac{\log N}{N}. \tag{1}$$

This bound was presented as a sum of two terms: The first term, henceforth referred to as *parameter redundancy* (PR), corresponds to universality w.r.t. the unknown source parameters within each stationary segment. It consists of $0.5 \log N/N$ bits per symbol for each component of the parameter vector and each stationary segment (see, Rissanen [13]). The second term, henceforth referred to as *transition redundancy* (TR), corresponds to universality w.r.t. the unknown transition times from one stationary segment to another. This term consists of $\log N/N$ bits per symbol for each such transition.

In [11], Merhav also demonstrated a universal compression scheme for the PSMS that achieves this lower bound. This scheme employs the method of mixtures in two stages. The first-stage mixture gives Krichevsky-Trofimov probability estimates [10] for each set of transition times, henceforth referred to as *transition path*. The second-stage mixture is performed over all possible transition paths. A *strongly sequential* version of this scheme was also obtained. That is, a scheme that sequentially updates the conditional coding probability of the next symbol given the past, independently of the future and of the horizon N. Merhav's scheme is of linearly increasing per-letter coding complexity, when we assume at most a single transition. It can be generalized to any fixed number of transitions, yielding an algorithm of polynomially increasing complexity, and to an exponentially increasing complexity scheme if no assumption of a fixed number of transitions is made.

Three strongly sequential schemes of smaller, but still increasing complexity, for universal coding of PSMS's were later proposed by Willems [16]-[19]. These schemes are all based on context tree coding [20] combined with arithmetic coding. They all obtain redundancy of at least $O(\log N/N)$, but with coefficients larger than the coefficient of the lower bound in [11]. Willems implemented two-stage mixtures as described above by constructing suitable state diagrams for the second-stage mixture. The weight of a transition path in the mixture is hence obtained by state transition weights along the path. The first two schemes ([16]-[18]) take into account all transition paths. In [17] and [18] all transition paths that assume the same most recent transition time are unified into one trellis state. This results in linearly increasing per-letter computational complexity, storage complexity of $O(N \log N)$ and redundancy of $0.5(C+1) \log N/N$ beyond the lower bound. The second scheme [16], [17] groups transition paths into states according to both the last transition time and the hypothesized number of transitions thus far. The diagram results in more states, each representing less transition paths. This leads to quadratically increasing per-letter complexity and a total redundancy of $0.5 \log N/N$ beyond the bound. The third approach, proposed recently in [19], selectively eliminates states according to the time they were created. This scheme is still of

increasing complexity of $O(\log N)$ and achieves per-letter redundancy of $O\left(\log^2 N/N\right)$ overall. To the best of our knowledge, no fixed complexity universal scheme has been obtained for PSMS's.

In this paper, we derive, analyze and present simulation results of three new universal strongly sequential data compression schemes for PSMS's based on context tree coding. The first scheme achieves the lower bound on the redundancy, whereas the two other schemes provide slower decay rate of the redundancy, but have *fixed complexity* (and logarithmic bit storage complexity). The first scheme is a generalization of Merhav's scheme which uses Willems' linear transition diagram with different weights. Similarly as Willems' scheme, it is of linearly increasing per-letter complexity. Unlike all the schemes presented in [11], [16]-[19], for which the number of states grows with time, the last two schemes have fixed number of states, and hence can be applied for practical purposes. Although the convergence rate does not meet the bound, it is better than any existing low-complexity scheme for PSMS's.

The second scheme combines decisions based on the observed past with a reduced-state transition diagram, using different state transition weights than those used by Willems. It uses decisions to eliminate unlikely states in the diagram, thus preserving a fixed number of states. This scheme achieves *average* redundancy of $O(\log N/N)$ for large transitions if the number of stationary segments is upper bounded by some constant S, that depends on the design parameters of the scheme. Otherwise, *pointwise* redundancy (for any N-tuple) of at most $O(\log\log N/\log N)$ is obtained. In the third scheme, we partition the data sequence into smaller blocks and encode each one separately. Using the optimal block length, we achieve pointwise redundancy of $O\left(\sqrt{\log N/N}\right)$. Simulations show that the true redundancies of the last two schemes are even better than the upper bounds obtained.

We can easily combine both fixed complexity schemes to obtain the better redundancy between the two for any sequence. We hence obtain a maximum upper bound of $O\left(\sqrt{\log N/N}\right)$, which is better than the currently known $O(1/\log N)$ upper bound of the Lempel-Ziv algorithm.

The outline of this paper is as follows. Section 2 contains notation and definitions. In Section 3 we generalize the analysis for the redundancy of a coding scheme. Section 4 presents the first scheme of linear per-letter complexity. In Section 5, we describe and analyze the second scheme of decisions and weighting. Section 6 presents the block partitioning scheme along with the analysis of its rate of convergence. Numerical results are presented in Section 7. Finally, in Section 8, we present the summary and conclusions of this work.

2 Notation and Definitions

Let $\{P_\theta\}$ be a parametric family of memoryless stationary PMF's of vectors whose components take on values in a finite alphabet Σ of size r. The parameter θ designates the $r-1$ dimensional vector of letter probabilities. A string drawn by the source from time instant i to time instant j $(x_i, x_{i+1}, \ldots, x_j)$, $j > i$, will be denoted by x_i^j. Let $x_1^N \stackrel{\triangle}{=} x^N \stackrel{\triangle}{=} (x_1, x_2, \ldots, x_n, \ldots, x_N)$ be a string emitted from an r-ary PMF whose parameter θ takes on a particular value θ_0 from $n = 1$ to $n = t_1 - 1$; then $\theta = \theta_1$ from $n = t_1$ until $n = t_2 - 1$, and so on. Finally, from $n = t_C$ to $n = N$, θ is held at θ_C. The vectors $\{x_1, \ldots, x_{t_1-1}\}$, $\{x_{t_1}, \ldots, x_{t_2-1}\}, \cdots, \{x_{t_C}, \ldots, x_N\}$ will be referred to as *stationary segments*, and correspondingly, $\theta_0, \theta_1, \cdots, \theta_C$ will be called the *segmental parameters*. It will be assumed that the different segments are statistically independent. The extended vector $(\theta_0, \theta_1, \ldots, \theta_C)$ will be denoted by Θ, and will be referred to as the *parameter set*. The C dimensional vector, representing the C time instants *before* which transitions take place, $(t_1, t_2, \ldots, t_C)$, will be denoted by $\mathcal{T}$, and referred to as the *true transition path*. For convenience, we define $t_0 \stackrel{\triangle}{=} 1$ and $t_{C+1} \stackrel{\triangle}{=} N + 1$. We will assume that the number of transitions C is either fixed or is of lower order than the time n. Noting that C is a function of the dimension of the other parameters, the PMF of the PSMS is parameterized by the pair $\{\Theta, \mathcal{T}\}$, and defined as follows.

$$P\left(x^N \mid \Theta, \mathcal{T}\right) = \prod_{i=0}^{C} P_{\theta_i}\left(x_{t_i}, \ldots, x_{t_{i+1}-1}\right), \tag{2}$$

where the PMF of each segment is obtained by

$$\begin{aligned} P_{\theta_i}\left(x_{t_i}, \ldots, x_{t_{i+1}-1}\right) &= \prod_{n=t_i}^{t_{i+1}-1} P_{\theta_i}\left(x_n\right) \\ &= \prod_{u \in \Sigma} P_{\theta_i}\left(u\right)^{n_{t_i}^{t_{i+1}-1}(u)}, \end{aligned} \tag{3}$$

where $P_{\theta_i}(x_n)$ is the probability of the letter x_n drawn by P_{θ_i}, which for simplicity, will be denoted by P_i, and $n_{t_i}^{t_{i+1}-1}(u)$ denotes the number of occurrences of $u \in \Sigma$ within the i-th segment.

The per-letter average entropy of a PSMS is obtained by

$$H\left(\Theta, \mathcal{T}\right) \stackrel{\triangle}{=} \frac{1}{N} \sum_{i=0}^{C} \left(t_{i+1} - t_i\right) H\left(\theta_i\right), \tag{4}$$

where $H(\theta_i)$ is the entropy of the i-th segment.

Since we assume no prior knowledge of $\{\Theta, \mathcal{T}\}$, we will not be able to assign the true probability of the sequence for a coding scheme. Instead, we will seek a universal sequential probability assignment that will implement a two stage mixture and will serve as the basis for arithmetic

coding. The probability assigned to the substring x^n by an algorithm $\mathcal{A}$ will be denoted by $Q_{\mathcal{A}}(x^n)$. To enable sequential updating of $Q_{\mathcal{A}}$, the conditional probability $Q_{\mathcal{A}}(x_n \mid x^{n-1})$ defined by [13] as

$$Q_{\mathcal{A}}\left(x_n \mid x^{n-1}\right) \stackrel{\triangle}{=} Q_{\mathcal{A}}(x^n) / Q_{\mathcal{A}}\left(x^{n-1}\right), \tag{5}$$

must be well defined. Additionally, in order to enable the use of arithmetic coding, the assigned probability must fulfill the terms described in [20]

$$Q_{\mathcal{A}}\left(x_1^{n-1}\right) = \sum_{x_n \in \Sigma} Q_{\mathcal{A}}(x_1^n), \ \forall x_1^{n-1} \in \Sigma^{n-1}, \tag{6}$$

where the probability of the empty string is one by convention.

The first-stage mixture, implemented to obtain the assigned probability, is performed for a given transition path. The conditional probability assignment $Q\left(x^N \mid \mathcal{T}'\right)$ given any transition path $\mathcal{T}' \stackrel{\triangle}{=} (t'_1, \ldots, t'_C)$, is recursively defined using the Krichevsky-Trofimov (KT) empirical estimates [10], that result from mixing the parameter with a Dirichlet(0.5) prior. The KT estimates will be used with relative frequency counts that are reset at every hypothesized transition. Specifically, the conditional letter probability is defined as

$$Q\left(X_t = u \mid x_1^{t-1}, \mathcal{T}'\right) \stackrel{\triangle}{=} \frac{n_{t'_i}^{t-1}(u) + \frac{1}{2}}{(n - t'_i) + \frac{r}{2}}, \quad t'_i \leq t < t'_{i+1}, \ \forall u \in \Sigma, \tag{7}$$

and the probability of an n-tuple is obtained by

$$\begin{aligned} Q(x_1^n \mid \mathcal{T}') &= \prod_{i=1}^{n} Q\left(x_i \mid x_1^{i-1}, \mathcal{T}'\right) \\ &= Q\left(x_1^{n-1} \mid \mathcal{T}'\right) \cdot Q\left(x_n \mid x_1^{n-1}, \mathcal{T}'\right), \end{aligned} \tag{8}$$

where x_1^0 represents the null string, whose probability is one by convention.

To implement the second stage mixture, the probability assigned to an N-tuple will be a weighted sum of conditional probability assignments given transition paths. Each probability assumes a different path $\mathcal{T}'$ from a set of paths $\{\mathcal{T}\}_{\mathcal{A}}$, selected by the algorithm $\mathcal{A}$. Each path will be weighted with some *weight function* $W_{\mathcal{A}}(\mathcal{T}')$.

$$Q_{\mathcal{A}}\left(x^N\right) = \sum_{\mathcal{T}' \in \{\mathcal{T}\}_{\mathcal{A}}} W_{\mathcal{A}}(\mathcal{T}') \, Q\left(x^N \mid \mathcal{T}'\right). \tag{9}$$

The weight function must be non-negative for all $\mathcal{T}'$ and satisfy

$$\sum_{\mathcal{T}' \in \{\mathcal{T}\}_{\mathcal{A}}} W_{\mathcal{A}}(\mathcal{T}') = 1. \tag{10}$$

The idea of the coding scheme is to construct a group $\{\mathcal{T}\}_{\mathcal{A}}$ that always includes the true transition set $\mathcal{T}$, or at least a good estimate $\hat{\mathcal{T}}$ of $\mathcal{T}$, that induces large values of $Q\left(x^N \mid \hat{\mathcal{T}}\right)$ and $W_{\mathcal{A}}\left(\hat{\mathcal{T}}\right)$. Such an estimate $\hat{\mathcal{T}}$ defines a triplet $\left\{\hat{C}, \hat{\Theta}, \hat{\mathcal{T}}\right\}$, where $\hat{\Theta}$ is defined by convex combinations of the true segmental parameters θ_i along each hypothesized segment. The weights of each convex combination are the relative durations of each θ_i within the hypothesized segment of $\hat{\mathcal{T}}$. For example, if $\mathcal{T} = \{\alpha N + 1\}$, $0 < \alpha < 1$, i.e., a single true transition, but we estimate no transitions, i.e., $\hat{\mathcal{T}} = \phi$, then $\hat{\Theta} = \left\{\hat{\theta}_0\right\}$, where $\hat{\theta}_0 = \alpha\theta_0 + (1-\alpha)\,\theta_1$. We define the *probability assignment for an estimated PSMS* as $Q\left(x^N \mid \hat{\Theta}, \hat{\mathcal{T}}\right)$, similarly as in eq. (2).

3 The Redundancy

The *pointwise redundancy* of scheme $\mathcal{A}$ for an N-tuple x^N is defined as

$$R\left(x^N; \mathcal{A}\right) \stackrel{\triangle}{=} \frac{1}{N} \log \frac{P\left(x^N \mid \Theta, \mathcal{T}\right)}{Q_{\mathcal{A}}\left(x^N\right)}, \tag{11}$$

ignoring negligible integer length constraints. The (expected) N-th order *redundancy* of scheme $\mathcal{A}$ is defined as

$$R_N\left(\mathcal{A}\right) \stackrel{\triangle}{=} E_{\{\Theta,\mathcal{T}\}}\left[R\left(x^N; \mathcal{A}\right)\right], \tag{12}$$

where $E_{\{\Theta,\mathcal{T}\}}$ denotes the expectation w.r.t. a given PSMS $\{\Theta, \mathcal{T}\}$.

The pointwise redundancy of an N-tuple for a PSMS can be expressed as

$$\begin{aligned} R\left(x^N; \mathcal{A}\right) &= \frac{1}{N} \log \frac{Q\left(x^N \mid \hat{\mathcal{T}}\right)}{Q_{\mathcal{A}}\left(x^N\right)} + \frac{1}{N} \log \frac{Q\left(x^N \mid \hat{\Theta}, \hat{\mathcal{T}}\right)}{Q\left(x^N \mid \hat{\mathcal{T}}\right)} + \frac{1}{N} \log \frac{P\left(x^N \mid \Theta, \mathcal{T}\right)}{Q\left(x^N \mid \hat{\Theta}, \hat{\mathcal{T}}\right)} \\ &\stackrel{\triangle}{=} R_t\left(x^N; \mathcal{A}\right) + R_p\left(x^N; \mathcal{A}\right) + R_d\left(x^N; \mathcal{A}\right) \end{aligned} \tag{13}$$

Eq. (13) decomposes the pointwise redundancy into three terms: R_t - *transition redundancy* (TR), R_p - *parameter redundancy* (PR), and R_d - *decision redundancy* (DR). The TR reflects universality w.r.t. the transition path. The PR is the cost of universality w.r.t. Θ for a given transition set. The DR is the additional redundancy caused by estimation error of the transition path and the segmental parameters it imposes. These terms all depend on the specific algorithm, but of course, it is the total redundancy that should be compared to the lower bound.

Assigning the KT estimate to a stationary segment of length m results in additional $\frac{r-1}{2} \log m + O\left(1\right)$ code bits of unnormalized PR for that segment [10], [14], similarly to Rissanen's lower bound. Hence, using the conditional probability assignment of eqs. (7)-(8) for $Q\left(x^N \mid \hat{\mathcal{T}}\right)$ results in extra

code length, that is the sum of the additional code bits obtained for each of the segments imposed by $\hat{\mathcal{T}}$. Using the Jensen inequality and normalizing by N, we conclude that the PR is upper bounded by

$$R_p\left(x^N;\mathcal{A}\right) \leq \frac{r-1}{2}\left(\hat{C}+1\right)\frac{\log\frac{N}{\hat{C}+1}}{N} + O\left(\frac{\hat{C}}{N}\right) \tag{14}$$

From eq. (9) and the definition of TR in (13), we obtain that the TR depends on the weight of $\hat{\mathcal{T}}$ and is upper bounded by

$$\begin{aligned} R_t\left(x^N;\mathcal{A}\right) &\stackrel{\triangle}{=} \frac{1}{N}\log\frac{Q\left(x^N \mid \hat{\mathcal{T}}\right)}{Q_{\mathcal{A}}\left(x^N\right)} \\ &= \frac{1}{N}\log\frac{Q\left(x^N \mid \hat{\mathcal{T}}\right)}{\sum_{\mathcal{T}'\in\{\mathcal{T}\}_{\mathcal{A}}} W_{\mathcal{A}}\left(\mathcal{T}'\right) Q\left(x^N \mid \mathcal{T}'\right)} \\ &\leq -\frac{1}{N}\log W_{\mathcal{A}}\left(\hat{\mathcal{T}}\right). \end{aligned} \tag{15}$$

The inequality holds by definition of $\hat{\mathcal{T}}$. The DR results from coding more than one true stationary segment as if it were a single stationary block. Assume the block x_{i+1}^{i+m} of length m contains data drawn by s distributions P_{u+1} to P_{u+s}, and assume this block is coded as if it were drawn by a stationary PMF Q, then Q is defined as the convex combination of the true distributions, where each PMF P_{u+l} is weighed by its relative duration in the block. It is easy to show that the contribution to the DR of this block is upper bounded by the entropy of the relative durations vector multiplied by m and normalized by the complete sequence length N. Since the vector consists of s components, we can bound the contribution of this block to the DR by

$$R_d\left(x_{i+1}^{i+m};\mathcal{A}\right) \leq \frac{m\log s}{N}. \tag{16}$$

The DR of an N-tuple is the sum of the contributions of all segments hypothesized by $\hat{\mathcal{T}}$.

4 An Optimal Linear Per-Letter Complexity Scheme

The scheme presented in this section and denoted by $\mathcal{W}$, uses Willems' linear weighting scheme [17], [18] to group transition paths into states in a diagram, but with different weight functions, corresponding to the scheme in [11]. It will be shown to achieve the lower bound on the redundancy, as opposed to Willems' scheme.

The idea of Willems' linear scheme is to implement the mixture method using a *linear transition diagram* that contains all possible transition paths, as illustrated in Figure 1. This trellis reduces the exponential complexity and still enables weighting of all transitions paths. A directed path

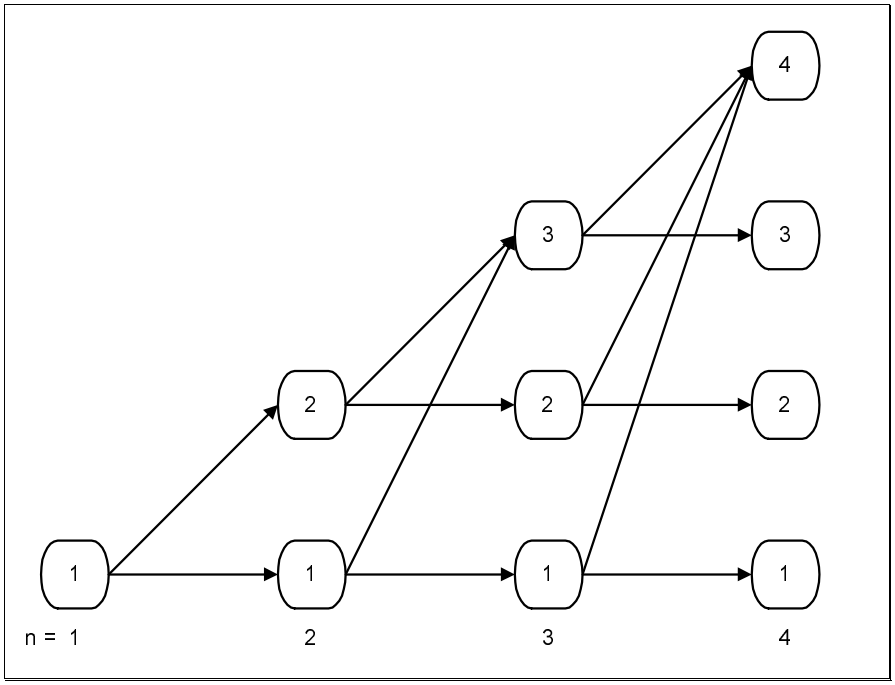

Figure 1: linear transition diagram at time instants 1 to 4. A number in a state box denotes the most recent transition point of the state represented by the box. The time is denoted below the graph.

along the diagram represents a transition path. Horizontal move denotes that the source remains in the same stationary segment. An upward move in the graph represents a transition of the source. A box in the diagram represents a state. State s_n at time n is defined as the time instant of the most recent transition within the period $1 \leq t \leq n$. Each state is assigned a weight $G\left(x_1^n \mid s_n\right)$ associated with the subsequence x_1^n, in order to implement the weighting procedure. The probability assigned to x_1^n is the sum of the weights of all states in the diagram at time n.

$$Q_{\mathcal{W}}\left(x_1^n\right) \stackrel{\triangle}{=} \sum_{s_n=1}^{n} G\left(x_1^n \mid s_n\right). \tag{17}$$

We note that by definition of a state, the diagram will always consist of n states at time n, that form a partition of all transitions paths into n disjoint sets.

The weight of a state is recursively defined by the KT estimates and the transition rules of the trellis. The KT probability of letter x_n at state s_n is obtained, similarly as in (7), by

$$Q\left(x_n \mid s_n\right) \stackrel{\triangle}{=} \frac{n_{s_n}^{n-1}\left(x_n\right) + \frac{1}{2}}{\left(n - s_n\right) + \frac{r}{2}}. \tag{18}$$

The only two possible transitions from state s_n at time n to state s_{n+1} at time $n+1$ are the self transition, i.e., $s_{n+1} = s_n$ and $s_{n+1} = n+1$, i.e., transition to the *only* new state formed at time $n+1$ that assumes a source transition between time n and time $n+1$. Associated with each such

transition, there is a weight $W_{tr}\left(s_{n+1} \mid s_n\right)$, where

$$W_{tr}\left(s_{n+1}=s_n \mid s_n\right)+W_{tr}\left(s_{n+1}=n+1 \mid s_n\right)=1. \tag{19}$$

The weight of a state is therefore recursively updated by

$$G\left(x_1^n \mid s_n\right)=Q\left(x_n \mid s_n\right)\cdot\begin{cases} W_{tr}\left(s_n=s_{n-1} \mid s_{n-1}\right)\cdot G\left(x_1^{n-1} \mid s_{n-1}\right), & s_n<n \\ \sum_{j=1}^{n-1} W_{tr}\left(s_n=n \mid s_{n-1}=j\right) G\left(x_1^{n-1} \mid s_{n-1}=j\right), & s_n=n \end{cases}. \tag{20}$$

It is easy to see that the weight assigned to a state by this procedure is a weighted sum of the KT estimates assigned to all transition paths $\mathcal{T}_n'$ of order n that lead to this state,

$$G\left(x_1^n \mid s_n\right)=\sum_{\mathcal{T}_n'\to s_n} W\left(\mathcal{T}_n'\right) Q\left(x_1^n \mid \mathcal{T}_n'\right), \tag{21}$$

where $\mathcal{T}_n'\to s_n$ is the set of all paths leading to s_n. The weight $W\left(\mathcal{T}_n'\right)$ is the product of all the transition weights along the path representing $\mathcal{T}_n'$ in the diagram.

So far, we have described a general linear scheme as presented in [17], [18]. The transition weights defined by Willems use the binary KT estimates of the distance from the last transition.

$$W_{tr}\left(s_{n+1} \mid s_n\right)\triangleq\begin{cases} \frac{\frac{1}{2}}{(n-s_n)+1}, & s_{n+1}=n+1 \\ \frac{(n-s_n)+\frac{1}{2}}{(n-s_n)+1}, & s_{n+1}=s_n \end{cases}. \tag{22}$$

The proposed scheme defines state transition weights $W_{tr}\left(s_{n+1} \mid s_n\right)$, that are different from those defined above. The weights are defined as follows. For a given $\varepsilon>0$, let:

$$\pi(j) \triangleq \frac{1}{j^{1+\varepsilon}}, \quad 1\le j<N; \tag{23}$$

$$C_n \triangleq \sum_{j=1}^{n}\pi(j), \tag{24}$$

$$\text{and } C_\infty = \sum_{j=1}^{\infty}\pi(j). \tag{25}$$

Now,

$$W_{tr}\left(s_{n+1} \mid s_n\right)\triangleq\begin{cases} \frac{\pi(n)}{C_\infty-C_{n-1}}, & s_{n+1}=n+1 \\ \frac{C_\infty-C_n}{C_\infty-C_{n-1}}, & s_{n+1}=s_n \end{cases}. \tag{26}$$

By eqs. (23)-(26), we assign to each time $n,\ 1\le n<N$, a distribution over the fixed time $t,\ t>n$ for the probability that the next transition occurs just before time t. The probability of source transition at time $n+1$ is the weight $W_{tr}\left(n+1 \mid s_n\right)$, which is selected to decay as $n^{-(1+\varepsilon)}$, causing additional TR of $(1+\varepsilon)\log n/N$ if a true transition occurs at that point. We note that the above weights depend on the absolute time, instead of the relative time from the last transition as in

(22). This reduces the weight of a transition, and is justified by the assumption that the number of transitions is small. This also leads to a computational improvement. First, the self-transition weight can be calculated *once* using (26) for all self-transitions in eq. (20). Furthermore, eq. (20) for $s_n = n$ reduces, using eq. (17), to

$$G\left(x_1^n \mid s_n = n\right) = Q\left(x_n \mid s_n\right) \cdot W_{tr}\left(s_n = n \mid s_{n-1} < n\right) Q_{\mathcal{W}}\left(x_1^{n-1}\right), \tag{27}$$

where $W_{tr}\left(s_n = n \mid s_{n-1} < n\right)$ is the same for all $s_{n-1} < n$. The drawback of the scheme with these weights is that it is computationally more demanding than the scheme with Willems' weights.

Willems' weights attain pointwise redundancy higher than the lower bound, presented in (1),

$$R\left(x^N; \mathcal{L}\right) \leq \left[\frac{r-1}{2}\left(C+1\right) + \frac{3C+1}{2}\right] \frac{\log N}{N} + O\left(\frac{C}{N}\right), \tag{28}$$

where $\mathcal{L}$ denotes Willems' scheme. The quadratical per-letter complexity scheme does not attain the bound either. The scheme with the proposed weights, on the other hand, attains the lower bound. This is stated in the following theorem.

Theorem 1 *The redundancy of the linear weighting scheme with state transition weights as in (26), is upper bounded by*

$$R\left(x^N; \mathcal{W}\right) \leq \left[\frac{r-1}{2}\left(C+1\right) + C + \varepsilon\right] \frac{\log N}{N} + O\left(\frac{C}{N}\right) \tag{29}$$

for every N-tuple drawn by any PSMS, for all $\varepsilon > 0$.

If we let ε decay *slowly* with time, such that $\varepsilon_j \stackrel{\triangle}{=} \frac{\log(\log 2j)^k}{\log j}$, $k > 1$, and

$$\pi\left(j\right) \stackrel{\triangle}{=} j^{-(1+\varepsilon_j)} = \frac{1}{j\left(\log 2j\right)^k}, \tag{30}$$

we have

$$R\left(x^N; \mathcal{W}\right) \leq \left[\frac{r-1}{2}\left(C+1\right) + C\right] \frac{\log N}{N} + O\left(\frac{C \log\log N}{N}\right). \tag{31}$$

We conclude this section with the proof of the theorem.

Proof of Theorem 1: It is straightforward that this coding scheme satisfies eq. (9) (see e.g. [17]), and weighs *all* possible transition paths. Therefore, the true path $\mathcal{T}$ is always included, and so, $R_d = 0$ and $\hat{\mathcal{T}} = \mathcal{T}$. The PR is upper bounded by (14), with $\hat{C} = C$, and thus attains the lower bound, as expressed by the first term of (29). It is thus sufficient to show that the TR attains the lower bound as well. We will show that

$$R_t\left(x^N; \mathcal{W}\right) \leq \frac{1}{N}\left[\left(C+\varepsilon\right)\log N + \left(C+1\right)\left(1+\varepsilon\right) - C\log\varepsilon\right]. \tag{32}$$

We will first obtain $W_{\mathcal{W}}(\mathcal{T})$ using (26), and then upper and lower bound the partial sum $C_\infty - C_n$, $\forall n : 0 \leq n < N$, by approximating the sum by an integral. Finally, we will apply these bounds to the cumulative weight function $W_{\mathcal{W}}(\mathcal{T})$ and upper bound TR by $-\log W_{\mathcal{W}}(\mathcal{T})/N$.

The weight of $\mathcal{T}$ is obtained by

$$\begin{aligned} W_{\mathcal{W}}(\mathcal{T}) &= \prod_{i=0}^{C}\left[W_{tr}\left(s_{t_i}=t_i \mid s_{t_i-1}=t_{i-1}\right)\cdot\prod_{j=t_i+1}^{t_{i+1}-1}W_{tr}\left(s_j=t_i \mid s_{j-1}=t_i\right)\right] \\ &= \prod_{n=1}^{N-1}W_{tr}\left(s_{n+1}=s_n \mid s_n\right)\cdot\prod_{i=1}^{C}\frac{W_{tr}\left(s_{t_i}=t_i \mid t_{i-1}\right)}{W_{tr}\left(s_{t_i}=t_{i-1} \mid t_{i-1}\right)} \\ &= \prod_{n=1}^{N-1}\frac{C_\infty - C_n}{C_\infty - C_{n-1}}\cdot\prod_{i=1}^{C}\frac{\pi(t_i-1)}{C_\infty - C_{t_i-1}} \\ &= \frac{C_\infty - C_{N-1}}{C_\infty}\cdot\prod_{i=1}^{C}\frac{\pi(t_i-1)}{C_\infty - C_{t_i-1}}. \end{aligned} \tag{33}$$

We define $W_{tr}\left(s_1=1|s_0=t_{-1}\right) \stackrel{\triangle}{=} 1$. The first equality is obtained by taking the transition weights along $\mathcal{T}$. The second equality is obtained by multiplication and division by the self-transition weights at points of true source transitions. We use the telescopic property of the first product to obtain the last equality.

Approximating the infinite sum by an integral, we bound the partial sum $C_\infty - C_n$ by

$$\frac{1}{\varepsilon(n+1)^\varepsilon} \leq C_\infty - C_n \leq \frac{2^{1+\varepsilon}}{\varepsilon(n+1)^\varepsilon}, \quad \forall n \geq 0. \tag{34}$$

We note that $C_\infty = C_\infty - C_0$. Finally, we use this bound to upper bound the transition redundancy by the upper bound of eq. (15).

$$\begin{aligned} R_t\left(x^N;\mathcal{W}\right) &\leq -\frac{1}{N}\log W_{\mathcal{W}}(\mathcal{T}) \\ &= \frac{1}{N}\left[\sum_{i=1}^{C}\log\frac{C_\infty - C_{t_i-1}}{\pi(t_i-1)} - \log\left(C_\infty - C_{N-1}\right) + \log C_\infty\right] \\ &\leq \frac{1}{N}\left[\sum_{i=1}^{C}\log\frac{2^{1+\varepsilon}(t_i-1)^{1+\varepsilon}}{\varepsilon t_i^\varepsilon} + \log\left(\varepsilon N^\varepsilon\right) + \log\frac{2^{1+\varepsilon}}{\varepsilon}\right] \\ &\leq \frac{1}{N}\left[\sum_{i=1}^{C}\log t_i^{1+\varepsilon-\varepsilon} + \varepsilon\log N + (C+1)(1+\varepsilon) - C\log\varepsilon\right] \\ &\leq \frac{1}{N}\left[(C+\varepsilon)\log N + (C+1)(1+\varepsilon) - C\log\varepsilon\right]. \end{aligned} \tag{35}$$

The last inequality is obtained by taking N as an upper bound on t_i. This proves that the TR attains the lower bound, and therefore concludes the proof of the theorem.

5 A Decision Weighting Scheme

In this section we show that there exists a fixed complexity scheme, based on the transition diagram of the linear per-letter complexity scheme, that achieves vanishing redundancy. The redundancy is of the order of the lower bound when the transitions are large. We refer to the new scheme as the *decision weighting scheme* (DW), and denote it by $\mathcal{D}$. This scheme uses a *data-dependent reduced-state transition diagram*. It eliminates transition paths with low likelihood and it does not create a new state every time instant.

The scheme produces new states every $k \geq 1$ time instants instead of every instant, in order to reduce the diagram's growth rate. This forms a partition of the data into N/k non-overlapping blocks of length k, and for each block only one state is created. The parameter k is a design parameter, that will be referred to as the *block length*. In order to keep the number of *surviving* states (and the computational complexity) fixed, we assign to each state s a *metric* $M(s)$, that determines the likelihood of transition within the block represented by s. States with low metric values are eliminated. The number of high metric surviving states S is the second design parameter of the algorithm. By definition of the transition diagram, the set of surviving states defines a set of surviving transition paths, and a transition path that leads to an eliminated state is said to be eliminated and not to exist in the diagram.

The state number s represents the block number in which the most recent transition is assumed to have occurred. A state s, $s > 1$, that estimates transition at any time within the block it represents, is created at the block midpoint. The first state $s = 1$ is naturally created at the first time instant.

The metric of a state s is defined as follows. Let $n_s(u)$ be the number of occurrences of the the letter u in block s. The empirical per-letter entropy of the block is obtained by

$$H(s) = -\sum_{u \in \Sigma} \frac{n_s(u)}{k} \log \frac{n_s(u)}{k}. \tag{36}$$

The empirical entropy of the concatenation of blocks r and s is obtained by

$$H(r,s) = -\sum_{u \in \Sigma} \frac{n_r(u) + n_s(u)}{2k} \log \frac{n_r(u) + n_s(u)}{2k}. \tag{37}$$

Now, $M(s)$ is defined by

$$M(1) \stackrel{\triangle}{=} \infty \tag{38}$$

$$M(s) \stackrel{\triangle}{=} H(s-1, s+1) - \frac{1}{2}H(s-1) - \frac{1}{2}H(s+1); \quad \forall s > 1. \tag{39}$$

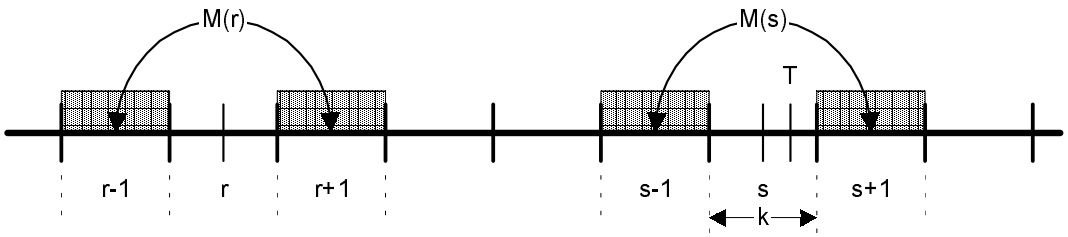

Figure 2: block partitioning for DW. The solid line represents the data sequence, that is divided into blocks of length k. True transition occurs at T and can be only estimated at the block midpoint by state s with likelihood $M(s)$, obtained by the empirical data of the two neighboring blocks $s-1$ and $s+1$. Another estimated transition is within block r estimated at its midpoint, with likelihood $M(r)$.

The quantity $M(s)$ measures the 'distance' between the empirical distributions of blocks $s-1$ and $s+1$. If $M(s)$ is large then it is likely that a change has occurred within block s. It is well-known that $M(s)$ serves as asymptotically optimal statistics for testing weather or not two sequences emerged from the same source [6], [23]. If $M(s)$ is large the scheme will assume transition which will be referred to the midpoint of block s by state s in the trellis. Figure 2 illustrates the partitioning mechanism. The metric $M(s)$ is non-negative for all $s > 1$, even if transition has not occurred. This can cause elimination of $s = 1$ if its metric had been defined smaller, even when other transitions have not occurred. Since state $s = 1$ always represents a transition, we thus define its metric to be infinite.

The probability assignment scheme can be described by the state diagram shown in Figure 3 for $k = 4$ and $S = 3$. The diagram begins when all S high metric states already exist, i.e., in steady state. The boxes in the diagram denote the states, and the numbers in the boxes the block numbers of the most recent transitions assumed by the states. As in the linear scheme, each state is assigned a weight $G(x_1^n \mid s_n)$ associated with the subsequence x_1^n. The weight of a state is recursively defined by the KT estimates and the transition rules shown in the diagram. The KT probability of letter x_n at state s_n is obtained by

$$Q(x_n \mid s_n) \stackrel{\triangle}{=} \frac{n_{\tau_n}^{n-1}(x_n) + \frac{1}{2}}{(n-\tau_n) + \frac{r}{2}}, \tag{40}$$

where the time τ_n is the first time instant after the last transition assumed by s_n, which is defined by

$$\tau_n \stackrel{\triangle}{=} \begin{cases} 1, & s_n = 1 \\ \lfloor (s_n - 0.5)\, k \rfloor + 1, & s_n > 1 \end{cases}. \tag{41}$$

The transition rules and update procedures at each time point are defined for three different cases as follows:

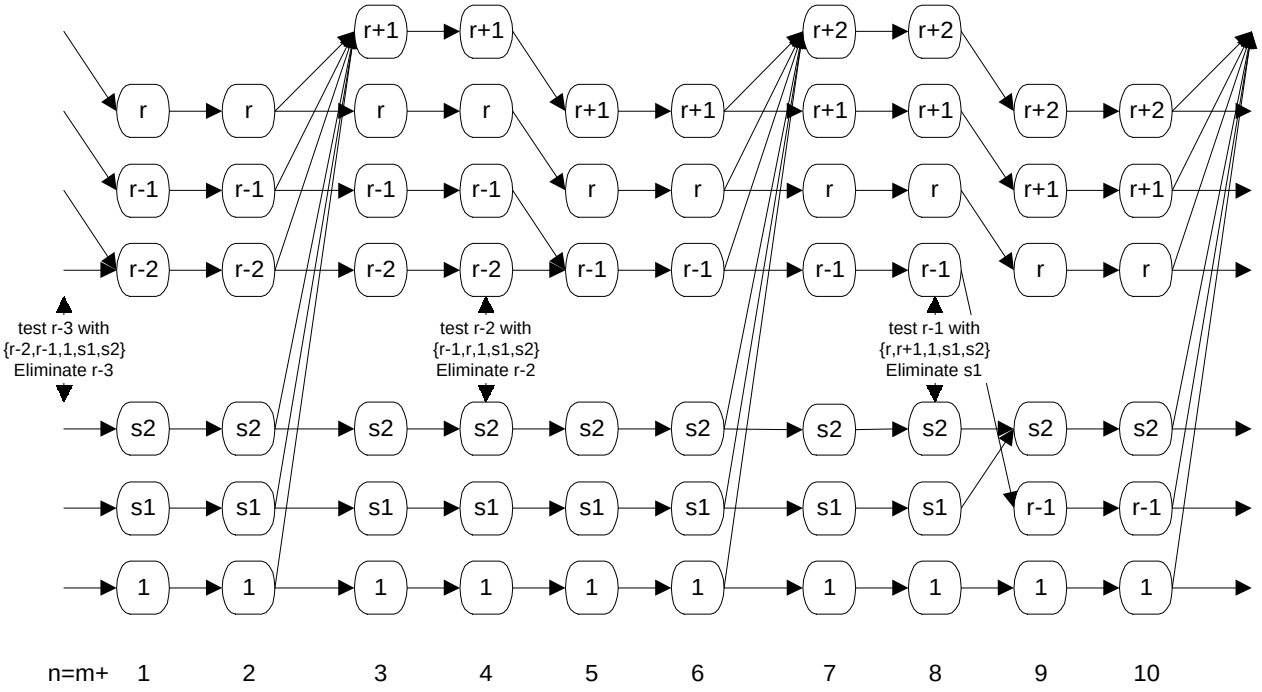

Figure 3: example of DW transition diagram in steady state for $k = 4$ and $S = 3$. The diagram starts right after time instant m at the first point of the $r + 1$ block. A single new state is created at any block midpoint, and a single low metric state is eliminated at any block partition point. Time instants are denoted below the diagram.

1. At a block midpoint $n = \lfloor (m + 0.5)\, k \rfloor + 1;\ m > 0$ fixed, a new state r is created. The state weights are recursively updated almost as in the linear scheme by

$$G\left(x_1^n \mid s_n\right) = Q\left(x_n \mid s_n\right) \cdot \begin{cases} W_{tr}\left(s_n = s_{n-1} \mid s_{n-1}\right) \cdot G\left(x_1^{n-1} \mid s_{n-1}\right), & s_n < r \\ \sum_{j=1}^{r-1} W_{tr}\left(s_n = r \mid j\right) G\left(x_1^{n-1} \mid s_{n-1} = j\right), & s_n = r \end{cases}, \quad (42)$$

   where the transition weights are obtained by

$$W_{tr}\left(s_{n+1} \mid s_n = j\right) \stackrel{\triangle}{=} \begin{cases} \frac{\pi(r-j)}{C_\infty - C_{r-j-1}}, & s_{n+1} = r \\ \frac{C_\infty - C_{r-j}}{C_\infty - C_{r-j-1}}, & s_{n+1} = j \end{cases}. \quad (43)$$

   (The weight of a previously eliminated state j is zero $G\left(x_1^{n-1} \mid j\right) = 0$). We note that the transition weights depend on the relative block number from the last transition instead of the absolute time, as in the linear scheme. The proof of Theorem 2 will be based on this definition.

2. At the first point of a new block $n = mk+1;\ m \geq 5$ fixed, at most a single state i is eliminated into another state j. The weight of i is added into the weight of j, which is the smallest state in the diagram that is still larger than i. Only the self-transition is performed from all other

states.

$$G\left(x_1^n \mid s_n\right) = Q\left(x_n \mid s_n\right) \cdot \begin{cases} 0, & s_n = i \\ G\left(x_1^{n-1} \mid i\right) + G\left(x_1^{n-1} \mid j\right), & s_n = j \\ G\left(x_1^{n-1} \mid s_n\right), & \text{otherwise} \end{cases} . \tag{44}$$

3. At any other point n there are only self-transitions.

$$G\left(x_1^n \mid s_n\right) = Q\left(x_n \mid s_n\right) \cdot G\left(x_1^{n-1} \mid s_n = s_{n-1}\right) \tag{45}$$

The elimination retains a fixed number of states in the diagram. It also ensures that no more than one state of any three consecutive states remain in the diagram. This is done to avoid a situation where a single transition is represented by two or three states. If a transition occurs at the midpoint of block s, all three states $s-1$, s and $s+1$ may have large metrics, and we only need to save one of them to represent the transition. We will therefore eliminate the states with the lower metric values among the three. The computation of the metric of state s requires delay of $1.5k$ time points, to obtain the empirical data of block $s+1$. An additional delay of $2k$ points is required for computation of the metrics of $s+1$ and $s+2$, that are tested against s. Hence, every new state will exist at least $3.5k$ time points before it is tested for the first time. Due to the delay of $3.5k$ time points between creation and the first possible elimination of a state, the steady state diagram contains $S+3$ states at time points of first halves of partition blocks and $S+4$ states at time points of second halves.

The elimination procedure at $n = mk+1$ takes three stages of testing the metric of the state s, created at $n-3.5k$. If $s-1$ or $s-2$ exist in the diagram, state s is eliminated, since $M\left(s-1\right) \geq M\left(s\right)$ or $M\left(s-2\right) \geq M\left(s\right)$. Otherwise, state s is tested against states $s+1$ and $s+2$, and if $M\left(s+1\right) > M\left(s\right)$ or $M\left(s+2\right) > M\left(s\right)$, s is eliminated. If s passed both tests and there are less than S states, created before $n-3.5k$, no elimination is performed. If there are S such states, the state with the lowest metric among the existing states, created at $n-3.5k$ or earlier, is eliminated. A state is always eliminated by adding its weight into the weight of the closest newer state. This strategy minimizes the DR in case a true transition is eliminated by replacing it by the closest hypothesized transition point, still existing in the diagram.

The probability assigned to the subsequence x^n is obtained, as in the linear scheme, by the sum of the weights of all states that exist in the diagram at time instant n.

$$Q_{\mathcal{D}}\left(x_1^n\right) = \sum_{s_n} G\left(x_1^n \mid s_n\right), \tag{46}$$

where the notation $\sum_{s_n}$ represents a sum over all states existing in the diagram at time instant n. In contrast to the linear scheme, this strategy does not satisfy the general mixture structure presented in eq. (9), but can be easily shown to yield a valid probability function.

The update procedure of the transition diagram is fully described in a flow-chart in Figure 4. The per-letter computational complexity of this scheme is $O(S)$. Since we will assume a fixed S, $O(S) = O(1)$. Every state stores occurrence counts of $O(N)$, therefore, the storage complexity of the scheme is $O(\log N)$.

We conclude this section with two theorems that upper bound the pointwise and average redundancies of the DW by expressions that vanish as k and N increase (as long as k is of smaller order). Both theorems show that the redundancy decays to zero, and specifically at the rate of the lower bound for large transitions. The next theorem upper bounds the pointwise redundancy.

Theorem 2 *The pointwise redundancy of the DW is bounded uniformly for all PSMS's by*

$$R\left(x^N;\mathcal{D}\right) \leq \frac{r-1}{6}\frac{\log k}{k} + O\left(\alpha\left(N,k\right)\right) \tag{47}$$

for every x^N*, where* $\alpha(N,k)$ *is defined by*

$$\alpha\left(N,k\right) \stackrel{\triangle}{=} \max\left\{\frac{1}{k}, \frac{Ck}{N}, \frac{\log N}{N}\right\}. \tag{48}$$

The proof of the theorem is presented in Appendix A. It is based on the choice of the transition weights in eq. (43), that depend on the relative block number from the last transition instead of the absolute time.

The DW scheme performs decisions. Obviously, there is a trade-off consideration associated with the choice of the parameter k. A larger k provides a more reliable metric, leading to smaller probability of eliminating the best $\hat{\mathcal{T}}$. On the other hand, larger k increases the DR caused by estimation of transitions at block midpoints. An upper bound on the average N-th order redundancy as a function of k can be obtained based on the analysis in Appendix B. By differentiating this bound w.r.t. k, it can be shown that the optimal choice of k is of the form

$$k = A\log N + O\left(\log\log N\right), \tag{49}$$

where the parameter A depends on the parameters of the PSMS. Since, we desire a universal scheme, we will define the block length as

$$k = A\log N, \tag{50}$$

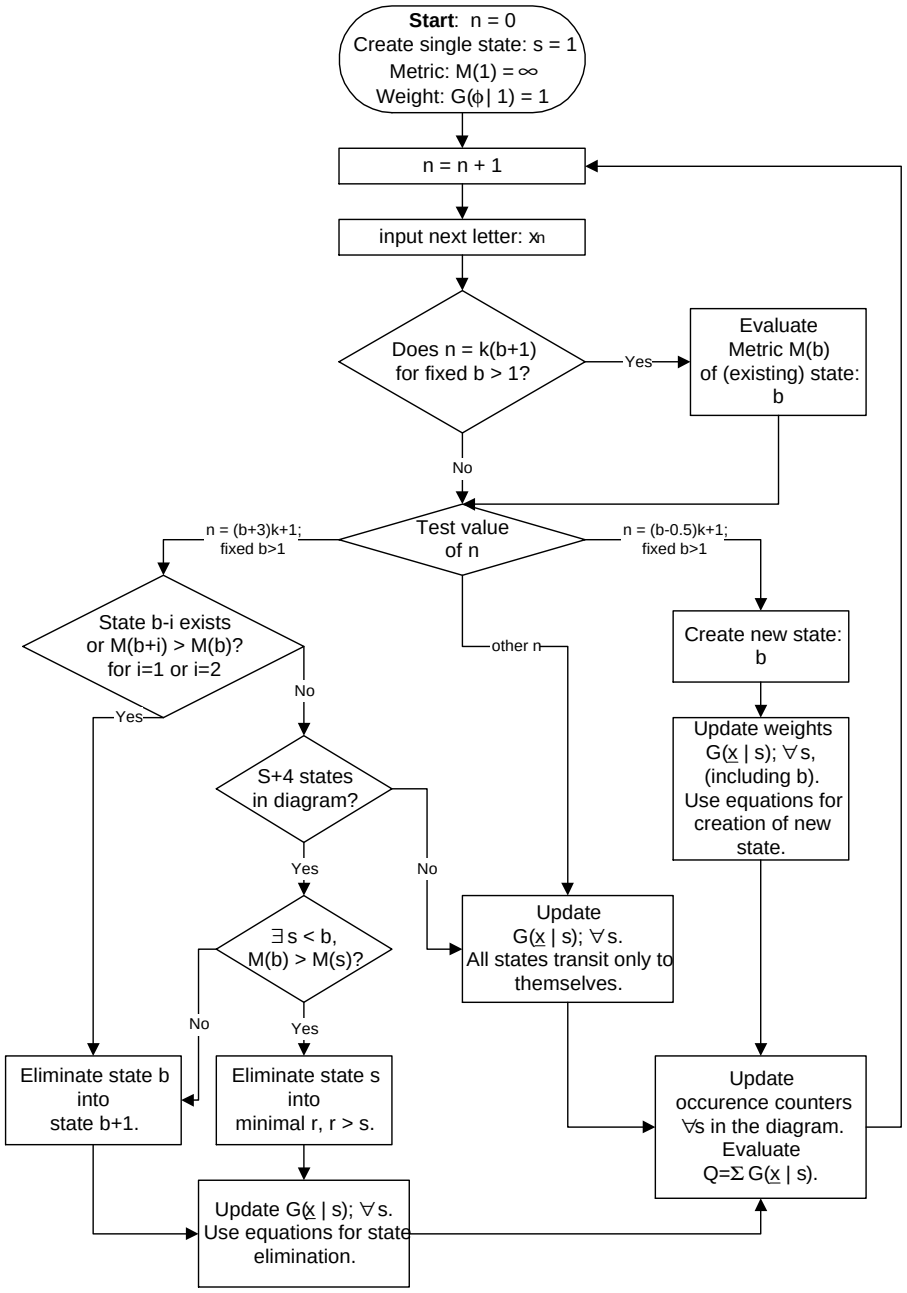

Figure 4: flow diagram of DW. Metrics of new states are evaluated at block end points. The diagram splits into the three different cases for updating the state weights. At block midpoints, new states are created. At block partition points, states are eliminated using the elimination criteria. At any other point, no transitions occur between different states.

where the parameter A will be a design parameter of the DW scheme. By substituting the block length of (50) in Theorem 2, we conclude that the pointwise redundancy for this choice of k is upper bounded by

$$R\left(x^N;\mathcal{D}\right) \leq \frac{r-1}{6A}\frac{\log\log N}{\log N} + O\left(\frac{1}{\log N}\right). \tag{51}$$

We now present the main theorem of the DW scheme. We begin with two definitions that characterize a PSMS. The *error exponent* of a PSMS, $E(\Theta)$, is defined as

$$E(\Theta) \stackrel{\triangle}{=} \min_{0\leq i\leq C-1}\left\{-2\log\frac{\sum_{u\in\Sigma}\sqrt{P_i(u)\,P_{i+1}(u)}+1}{2}\right\}. \tag{52}$$

The error exponent expresses the 'size' of the 'minimal' transition between adjacent segments of a PSMS. It can easily be shown that $0 \leq E(\Theta) \leq 2$. The larger is $E(\Theta)$, the larger is the 'minimal' transition of the PSMS.

The *divergence* (relative entropy) $D(P||Q)$ between distributions P and Q is defined as

$$D(P||Q) \stackrel{\triangle}{=} \sum_{u\in\Sigma} P(u)\log\frac{P(u)}{Q(u)}. \tag{53}$$

We define the *mean PSMS DR divergence* as

$$D(\Theta) \stackrel{\triangle}{=} \frac{1}{C}\sum_{i=1}^{C}\left\{D(P_{i-1}||P_i) + D(P_i||P_{i-1})\right\}. \tag{54}$$

Theorem 3 *Let $k = A\log N$. Assume a PSMS $\{\Theta,\mathcal{T}\}$ of at most S segments, each of length larger than $O(k)$. Then, the average N-th order redundancy of the DW scheme is upper bounded by*

$$R_N(\mathcal{D}) \leq \begin{cases} K_1\frac{\log N}{N} + o\left(\frac{C\log N}{N}\right), & \text{if } E(\Theta) > \frac{2}{A} \text{ and } D(\Theta) < \infty \\ 2.5AC\frac{\log^2 N}{N} + O\left(\frac{C\log N}{N}\right), & \text{if } E(\Theta) > \frac{2}{A} \text{ and } D(\Theta) = \infty \\ K_2\frac{\log^{r-1} N}{N^{AE(\Theta)-1}} + O\left(\frac{C\log N}{N}\right), & \text{if } \frac{1}{A} < E(\Theta) \leq \frac{2}{A} \\ \frac{r-1}{6A}\frac{\log\log N}{\log N} + O\left(\frac{1}{\log N}\right), & \text{if } E(\Theta) \leq \frac{1}{A} \end{cases}, \tag{55}$$

where

$$K_1 \stackrel{\triangle}{=} \left[\frac{r-1}{2}(C+1)\right] + \left[(1+\varepsilon)C+\varepsilon\right] + \left\{2.5A\left[\log e + D(\Theta)\right]C\right\}, \tag{56}$$

$$K_2 \stackrel{\triangle}{=} (A+1)^{r-1}C\log(C+1), \tag{57}$$

$O(\alpha)$ denotes the order of α, and $o(\alpha)$ denotes order smaller than the order of α.

The theorem demonstrates that we can achieve the order of the lower bound with a fixed complexity scheme if the transitions are large enough, while for smaller transitions we can still achieve decaying

redundancy. The proof of the theorem is presented in Appendix B. The bounds obtained are not tight. Tighter bounds of the same orders may be obtained by a much more complicated analysis than the one presented in Appendix B.

We note the different behavior of the average redundancy for different transition 'sizes'. If $E(\Theta) > 1/A$, it is likely that there exists a surviving path $\hat{\mathcal{T}}$, s.t. $\hat{C} = C$ and all transitions are estimated near their true locations. For large transitions, $E(\Theta) > 2/A$, the redundancy is mostly influenced by the block partitioning, i.e., by estimating transitions only at block midpoints. This factor increases if the PSMS contains transition of infinite divergence, thus obtaining the second region of the bound. When the transitions are smaller, $1/A < (\Theta) \leq 2/A$, the redundancy is determined by the probability that the smallest transition is not detected near its true location. If the PSMS contains very small transitions, $E(\Theta) \leq 1/A$, the scheme cannot ensure a good estimate of $\mathcal{T}$, and thus, we obtain redundancy of higher order.

The condition that bounds the number of segments ensures that if the last transition is the smallest one, the scheme will still create a surviving state for it. However, the condition restricting to segments larger than $O(k)$ is required only for mathematical convenience purposes. It obtains the very simple expressions for the error exponent and the PSMS divergence, presented above, but has no effect on the nature of the results. Therefore, we can obtain bounds of the same order for PSMS's with shorter segments, but the mathematical representations of the error exponent and the PSMS divergence will become much more complex, and dependent on $\mathcal{T}$. (Viewing the first region of the upper bound, shorter segments will increase the low order term to be of the order of the first term).

6 Block Codes

The DW scheme achieves low order redundancy for some PSMS's, while for others it achieves redundancy that vanishes very slowly as $O(\log\log N/\log N)$. It can be shown that the LZ-78 algorithm has a pointwise upper bound with the same rate. We desire a scheme that attains better redundancy for the second group of sources, for which the DW scheme performs poorly and does not achieve better rate than LZ. In this section we present such a scheme, that can be combined with the DW scheme in order to achieve the better redundancy for any PSMS. The new scheme is referred to as the *block partitioning* (BP) scheme and will be denoted by $\mathcal{P}$.

The BP scheme partitions the N-tuple into B blocks, and codes each block b, $1 \leq b \leq B$ of

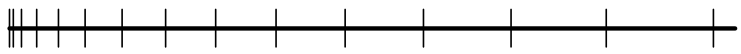

Figure 5: description of block partitioning. The horizontal line represents the time axis and the vertical lines the partition points.

length m_b as if it were a stationary segment, using its KT estimate. The probability assigned to an N-tuple is defined as

$$Q_{\mathcal{P}}\left(x^N\right) \stackrel{\triangle}{=} Q\left(x^N \mid \hat{\mathcal{T}}\right), \tag{58}$$

where $\hat{\mathcal{T}} \stackrel{\triangle}{=} \left\{\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_{B-1}\right\}$ is independent of x^N and is recursively defined as

$$\hat{t}_b \stackrel{\triangle}{=} \hat{t}_{b-1} + m_b, \;\; 1 \leq b \leq B-1, \tag{59}$$

where $\hat{t}_0 \stackrel{\triangle}{=} 1$. (Hence, by definition, $B \stackrel{\triangle}{=} \hat{C}+1$). The idea is to choose the set $\{m_b\}$ that will give the fastest decay of the *pointwise* redundancy *uniformly* over all PSMS's. We will achieve decay rate slower than $O(\log N/N)$ but faster than $O(\log\log N/\log N)$.

From eq. (15) we obtain that there is no TR, since there is a single transition path and no weighting. The redundancy is therefore obtained by trading off the PR, which decreases with the block length since $\hat{C}$ decreases (eq. (14)), and the DR which increases with the block length, (eq. (16)). It can be shown that for a given N, the best trade off is achieved by selecting the block length as

$$m_{opt} = O\left(\sqrt{N\log N}\right), \;\; \forall b. \tag{60}$$

Since N is unknown in advance, we can define the block length to increase with n, s.t. at time instant n the length of a block will be $O\left(\sqrt{n\log n}\right)$. The BP block length, obtained by the following equation, satisfies this requirement.

$$m_b = \lfloor kb\log b \rfloor \,; \;\; b > 1, \tag{61}$$

where $m_1 \stackrel{\triangle}{=} 1$. The parameter k is a design parameter. This assignment ensures that the last blocks will be $O\left(\sqrt{N\log N}\right)$ and larger than the preceding blocks, and therefore will dominate the redundancy. Figure 5 demonstrate the partitioning of an N-tuple.

Theorem 4 *The pointwise redundancy of the BP is bounded uniformly for all PSMS's with C transitions by*

$$R\left(x^N;\mathcal{P}\right) \leq \left(\frac{r-1}{2\sqrt{k}} + C\sqrt{k}\right)\sqrt{\frac{\log N}{N}} + o\left(\sqrt{\frac{\log N}{N}}\right); \quad \forall x^N. \tag{62}$$

The proof of Theorem 4 is presented in Appendix E. It is easy to show that if C is known in advance, the choice of

$$k = \frac{r-1}{2C} \tag{63}$$

will obtain the best upper bound

$$R\left(x^N;\mathcal{P}\right) \leq \sqrt{2\left(r-1\right)C}\sqrt{\frac{\log N}{N}} + o\left(\sqrt{\frac{\log N}{N}}\right); \quad \forall x^N. \tag{64}$$

If C is unknown, we can choose $k = \left(r-1\right)/2$ to obtain

$$R\left(x^N;\mathcal{P}\right) \leq \frac{\left(C+1\right)\sqrt{r-1}}{\sqrt{2}}\sqrt{\frac{\log N}{N}} + o\left(\sqrt{\frac{\log N}{N}}\right); \quad \forall x^N. \tag{65}$$

The BP scheme is very simple to implement and requires a single state only. Its per-letter computational complexity is $O\left(1\right)$ and storage complexity $O\left(\log N\right)$.

We can easily combine the DW and the BP schemes into a combined scheme, denoted by $\mathcal{C}$. Obviously, the probability assignment

$$Q_{\mathcal{C}}\left(x^N\right) \stackrel{\triangle}{=} \frac{1}{2}Q_{\mathcal{D}}\left(x^N\right) + \frac{1}{2}Q_{\mathcal{P}}\left(x^N\right) \tag{66}$$

attains the minimal redundancy between the two schemes. The pointwise redundancy of this scheme is always upper bounded by $O\left(\sqrt{\frac{\log N}{N}}\right)$ as in Theorem 4. If a PSMS satisfies the conditions of Theorem 3, its average N-th order redundancy is upper bounded by

$$R_N\left(\mathcal{C}\right) \leq \begin{cases} K_1\frac{\log N}{N} + o\left(\frac{C\log N}{N}\right), & \text{if } E\left(\Theta\right) > \frac{2}{A} \text{ and } D\left(\Theta\right) < \infty \\ 2.5AC\frac{\log^2 N}{N} + O\left(\frac{C\log N}{N}\right), & \text{if } E\left(\Theta\right) > \frac{2}{A} \text{ and } D\left(\Theta\right) = \infty \\ \left(\frac{r-1}{2\sqrt{k}} + C\sqrt{k}\right)\sqrt{\frac{\log N}{N}} + o\left(\sqrt{\frac{\log N}{N}}\right), & \text{otherwise} \end{cases}. \tag{67}$$

7 Simulation Results

In this section we present numerical examples of the performance of the schemes presented in Sections 4 through 6, and compare them to the performance of the schemes presented in [17]-[19]. We show that we achieve better performance with the new schemes and that the true redundancies are much smaller than the upper bounds.

Figure 6 compares the redundancies of both linear schemes (Willems' scheme and the optimal one) for a sequence of length 1000, drawn by a PSMS of $C = 3$ transitions. We use the optimal scheme with $\varepsilon = 0.1$ to obtain better redundancy than Willems' scheme. The true performance of both schemes is better than the upper bound of Theorem 1.

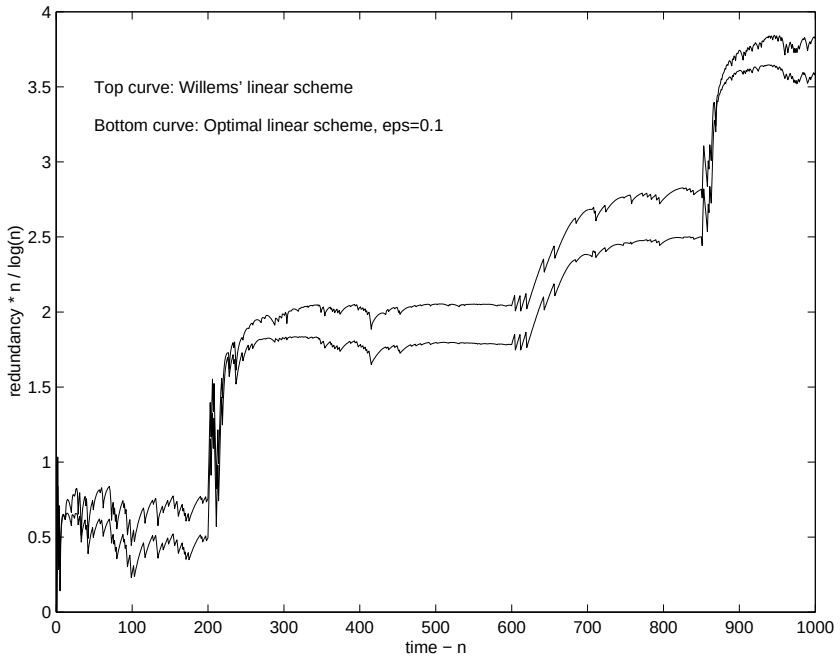

Figure 6: pointwise redundancies of linear schemes for $x_1^N, N = 10^3$ drawn by a binary PSMS, $\Theta = \{0.8, 0.2, 0.1, 0.4\}$, $\mathcal{T} = \{201, 601, 851\}$. The redundancies are multiplied by $n/\log n$. The redundancy of Willems' linear scheme is described by the top curve, and of the optimal scheme by the bottom curve.

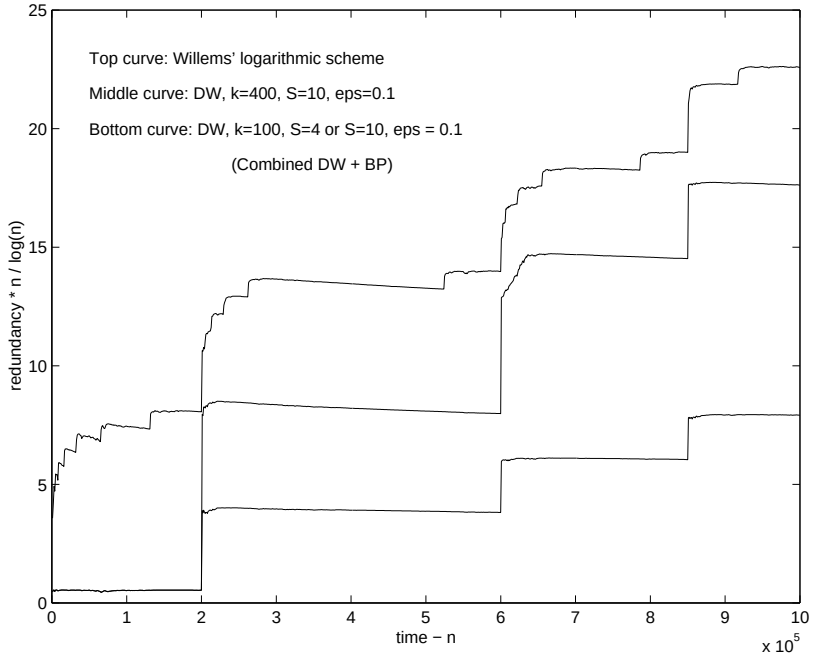

Figure 7: pointwise redundancies of the logarithmic scheme (top curve) and the DW scheme (bottom curves) for $x_1^N, N = 10^6$ drawn by a binary PSMS, $\Theta = \{0.8, 0.2, 0.7, 0.4\}$, $\mathcal{T} = \left\{2 \cdot 10^5 + 1, 6 \cdot 10^5 + 1, 8.5 \cdot 10^5 + 1\right\}$. The redundancies are multiplied by $n/\log n$. The parameters of the DW schemes are shown on the graph. The bottom curve describes DW scheme with $S = 4$ and with $S = 10$ and a combined DW BP scheme with the same DW parameters.

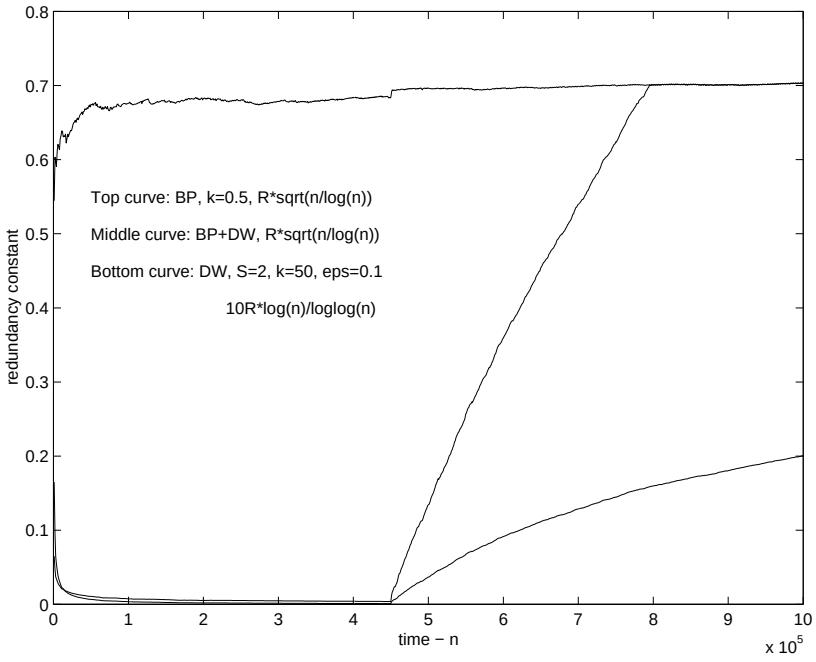

Figure 8: pointwise redundancies of the DW (bottom curve), BP (top curve) and the combined scheme (middle curve) for PSMS with a small transition. The N-tuple $x_1^N, N = 10^6$ is drawn by a binary PSMS, $\Theta = \{0.2, 0.1\}$, $\mathcal{T} = \left\{4.5 \cdot 10^5 + 1\right\}$. The redundancies of the upper two curves are multiplied by $\sqrt{n/\log n}$ and of the bottom curve by $10 \log n / \log\log n$. The parameters of the schemes are shown on the graph.

Figure 7 demonstrates pointwise redundancies obtained by Willems' logarithmic scheme, the DW scheme and the combined DW BP scheme for a binary PSMS with $C = 3$ large transitions. The DW scheme is shown to perform better than the logarithmic scheme. Figure 7 demonstrates that the DW scheme achieves redundancy $O(\log N/N)$, even for transitions for which $E(\Theta)$ is much smaller than $2/A$. For the PSMS in the example $E(\Theta) \approx 0.068$. Using block length $k = 100$, we have $2/A \approx 0.4$, but the DW still performs well. Using $k = 400$, we achieve higher redundancy, because transitions are estimated 400 points apart. The parameter S has no influence on the performance of the DW scheme as long as there are enough surviving states, which is the case for the bottom curve. Since the DW has better redundancy than the BP, the combined scheme obtains the redundancy of the DW scheme. Since $D(\Theta) \approx 1.52$, it is apparent that both curves representing the DW scheme attain much smaller redundancies than the upper bound of Theorem 3. This is because the bound is not tight.

Figure 8 illustrate the case of a single small transition, in which the BP performs better than the DW. The redundancy of the DW is $O(\log\log N/\log N)$. Before the transition occurs, the DW attains better redundancy. Hence, the combined scheme achieves this redundancy. After the transition occurred, the DW starts to perform poorly, until some point, where its redundancy

becomes higher than that of the BP. At that point, the combined scheme attains the redundancy of the BP. The redundancy of the BP is shown to be of $O\left(\sqrt{\log N/N}\right)$.

8 Summary and Conclusions

In this paper we investigated the problem of low complexity universal coding of a PSMS. We showed that the entropy of the source can be asymptotically achieved with fixed complexity schemes, and that these schemes can attain redundancies that decay faster than those obtained by any known low complexity scheme for coding PSMS's. Specifically, it was shown that the order of the lower bound can be achieved when the transitions in the statistics are large, and for smaller transitions the order of its square root is achieved. The lower bound itself was achieved by an optimal linear per-letter complexity scheme that was presented. Finally, all results were supported by simulations that showed that in practice all algorithms perform much better than the performance suggested by the analysis. All the schemes can be extended to more complex piecewise stationary sources using context tree coding schemes.

Appendix A – Proof of Theorem 2

We begin the proof of Theorem 2 with a lemma.

Lemma A.1 *Let $\hat{\mathcal{T}}$ be a transition path that is not eliminated from the transition diagram of the DW scheme, and let $\hat{C}$ be the number of transitions assumed by $\hat{\mathcal{T}}$. Then the pointwise TR is upper bounded as in (15) by*

$$\begin{aligned} R_t\left(x^N;\mathcal{D}\right) &\leq -\frac{1}{N}\log W_{\mathcal{D}}\left(\hat{\mathcal{T}}\right) \\ &\leq \frac{1}{N}\left[(1+\varepsilon)\,\hat{C}\log\frac{N}{k\hat{C}}+\varepsilon\log\frac{N}{k}+\left(\hat{C}+1\right)(1+\varepsilon)-\hat{C}\log\varepsilon\right], \end{aligned} \tag{A.1}$$

where $W_{\mathcal{D}}\left(\hat{\mathcal{T}}\right)$ is the cumulative weight assigned to the path $\hat{\mathcal{T}}$ by eq. (43).

Proof: The DW scheme weights all transition paths that survive in the diagram, with additional weights obtained from paths that lead to eliminated states. Hence, unlike equality (9),

$$\begin{aligned} Q_{\mathcal{D}}\left(x^N\right) &\geq \sum_{\mathcal{T}'\in\mathcal{D}} W_{\mathcal{D}}\left(\mathcal{T}'\right) Q\left(x^N \mid \mathcal{T}'\right) \\ &\geq W_{\mathcal{D}}\left(\hat{\mathcal{T}}\right) Q\left(x^N \mid \hat{\mathcal{T}}\right), \end{aligned} \tag{A.2}$$

where $\mathcal{T}' \in \mathcal{D}$ denotes the set of (surviving) transition paths that exist in the diagram at time N, and $\hat{\mathcal{T}}$ is one of these paths that is chosen as the estimate of the true transition path. Hence, by definition of TR in eq. (13), the first inequality of the lemma is proved, and we require an upper bound on $-\log W_{\mathcal{D}}\left(\hat{\mathcal{T}}\right)$ to prove the second. To attain this bound, we perform a similar procedure to the proof of Theorem 1. We begin with expressing $W_{\mathcal{D}}\left(\hat{\mathcal{T}}\right)$ as in eq. (33), as a product of the transition weights along the path. Since estimated transitions are fixed blocks apart, we count the partition block number instead of the time, and denote it by b, (where $\hat{b}_i$ represents the estimated block number of transition i, and s_j represents the state at time instant $n = \lfloor (j-0.5)\, k \rfloor + 1$ for $j > 1$, i.e. block midpoints, and $n = 1$ for $j = 1$).

$$
\begin{aligned}
W_{\mathcal{W}}\left(\hat{\mathcal{T}}\right) &= \left\{ \prod_{i=0}^{\hat{C}-1} \left[\prod_{j=\hat{b}_i+1}^{\hat{b}_{i+1}-1} W_{tr}\left(s_j = \hat{b}_i \mid s_{j-1} = \hat{b}_i\right) \right] \cdot W_{tr}\left(s_{\hat{b}_{i+1}} = \hat{b}_{i+1} \mid s_{\hat{b}_{i+1}-1} = \hat{b}_i\right) \right\} \cdot \\
&\quad \prod_{l=\hat{b}_{\hat{C}}+1}^{\lfloor N/k+0.5 \rfloor} W_{tr}\left(s_l = \hat{b}_{\hat{C}} \mid s_{l-1} = \hat{b}_{\hat{C}}\right) \qquad \text{(A.3)} \\
&= \prod_{i=0}^{\hat{C}-1} \left\{ \left[\prod_{j=\hat{b}_i+1}^{\hat{b}_{i+1}-1} \frac{C_\infty - C_{j-\hat{b}_i}}{C_\infty - C_{j-\hat{b}_i-1}} \right] \frac{\pi\left(\hat{b}_{i+1} - \hat{b}_i\right)}{C_\infty - C_{\hat{b}_{i+1}-\hat{b}_i-1}} \right\} \cdot \prod_{l=\hat{b}_{\hat{C}}+1}^{\lfloor N/k+0.5 \rfloor} \frac{C_\infty - C_{l-\hat{b}_{\hat{C}}}}{C_\infty - C_{l-\hat{b}_{\hat{C}}-1}} \\
&= \left[\prod_{i=1}^{\hat{C}} \frac{\pi\left(\hat{b}_i - \hat{b}_{i-1}\right)}{C_\infty} \right] \cdot \frac{C_\infty - C_{\lfloor N/k+0.5 \rfloor - \hat{b}_{\hat{C}}}}{C_\infty}.
\end{aligned}
$$

The first equality is obtained by assigning weight for the path in each of the $\hat{C}+1$ estimated segments. Each such weight is a product of self transitions within the segment multiplied by the transition weight to the next segment. The last term naturally does not exist for the last segment. We note that the number of points is $\lfloor N/k + 0.5 \rfloor$ instead of N since transitions only take place at block midpoints. The second equality is obtained by substituting the values of the weights using eq. (43). By the telescopic property of the products, we obtain the last equality. Taking the negative logarithm of the weight of the estimated transition path we attain

$$
-\log W_{\mathcal{D}}\left(\hat{\mathcal{T}}\right) = -\sum_{i=1}^{\hat{C}} \log \pi\left(\hat{b}_i - \hat{b}_{i-1}\right) + \left(\hat{C}+1\right) \log C_\infty - \log\left(C_\infty - C_{\lfloor N/k+0.5 \rfloor - \hat{b}_{\hat{C}}}\right). \qquad \text{(A.4)}
$$

The first term of the last equation can be upper bounded using Jensen inequality by

$$
\begin{aligned}
-\sum_{i=1}^{\hat{C}} \log \pi\left(\hat{b}_i - \hat{b}_{i-1}\right) &= \sum_{i=1}^{\hat{C}} (1+\varepsilon) \log\left(\hat{b}_i - \hat{b}_{i-1}\right) \qquad \text{(A.5)} \\
&\leq (1+\varepsilon)\, \hat{C} \log \frac{N}{k\hat{C}}.
\end{aligned}
$$

Using the last two equations, the bounds of (34) and the assumption that $\hat{b}_{\hat{C}} > 1$, we conclude the proof of the lemma.

Proof of Theorem 2: We now use the lemma to prove Theorem 2. The heart of the proof is the choice of $\hat{\mathcal{T}}$. Let $\hat{\mathcal{T}} \triangleq \{1.5k+1, 4.5k+1, 7.5k+1, \ldots\}$. The path $\hat{\mathcal{T}}$ is a partition of N into blocks of length $3k$. By definition of the DW scheme, a state is never eliminated before it is used for coding at least $3.5k$ data letters. Hence, the path $\hat{\mathcal{T}}$ always exists in the transition diagram and thus can be used to estimate $\mathcal{T}$. By definition of $\hat{\mathcal{T}}$, $N/(3k) - 1 \leq \hat{C} \leq N/(3k) + 1$. Using eq. (14), the PR can be upper bounded by

$$R_p\left(x^N; \mathcal{D}\right) \leq \frac{r-1}{6}\frac{\log k}{k} + O\left(\frac{1}{k}\right). \tag{A.6}$$

Substituting $\hat{C}$ in eq. (A.1), we bound the TR by

$$R_t\left(x^N; \mathcal{D}\right) \leq \frac{(1+\varepsilon)\log 6}{3k} + O\left(\frac{\log N}{N}\right) = \max\left\{O\left(\frac{1}{k}\right), O\left(\frac{\log N}{N}\right)\right\}. \tag{A.7}$$

To obtain the upper bound for the DR in the worst case where each transition contributes the most, we assume at most a single true transition in each block obtained by $\hat{\mathcal{T}}$. Since the blocks are of length $3k$, the DR is bounded, using (16), by

$$R_d\left(x^N; \mathcal{D}\right) \leq \frac{3Ck}{N} = O\left(\frac{Ck}{N}\right). \tag{A.8}$$

Since R_p is the dominant term, the theorem is proved.

Appendix B – Proof of Theorem 3

To prove Theorem 3, we address two different regions of the error exponent separately. For $E(\Theta) \leq 1/A$, we simply use the upper bound of Theorem 2 that applies to the redundancy of any sequence, and thus can be applied to the average redundancy for PSMS's with small error exponent.

We now prove the upper bounds for the other three regions. To analyze the average redundancy we select a surviving path $\hat{\mathcal{T}}$, that is most likely to be a good estimate of the true path $\mathcal{T}$. This path is used to form a division of all data sequences drawn by the PSMS into two disjoint sets. The first is the set of N-tuples for which $\hat{\mathcal{T}}$ is a good estimate, i.e., all transitions are estimated near their true time unit. The second set contains all the other N-tuples. We then upper bound the probability of each set and the average redundancies of all N-tuples in the set. Summing up both terms we obtain an upper bound for the average redundancy.

Let $\hat{\mathcal{T}} = \hat{\mathcal{T}}\left(x^N\right)$ be an estimate of the true path $\mathcal{T}$, s.t. $\hat{\mathcal{T}}$ connects the $C+1$ states of highest metric (including the first state $s=1$) and $\hat{C} = C$. It is assumed as a condition of the theorem that $C+1 \leq S$. Hence, we will always have $C+1$ states of highest metric, and therefore, this

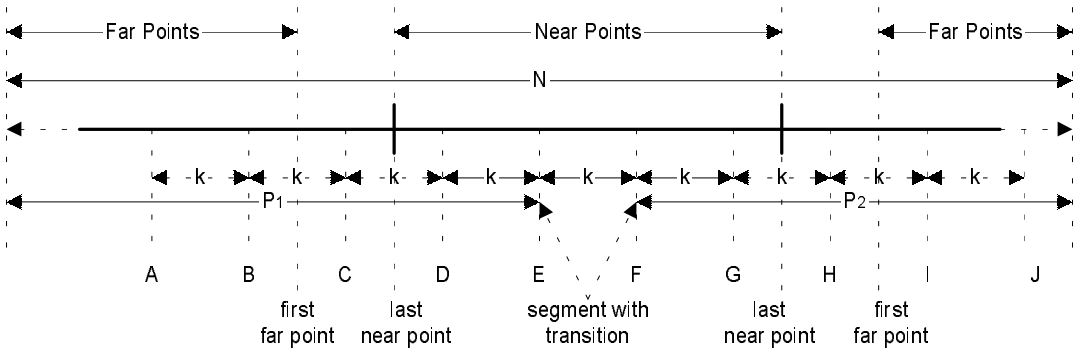

Figure B.1: far and near point definitions. The solid line represents the time axis. True transition occurs within the central segment. Points referred to as near points are at most $2.5k$ time units away from the transition. The first far point is one block away from the last near point, and obtains its metric by empirical data of blocks that do not overlap the blocks used to obtain the metric of the true transition.

path will exist. Each transition $t_i \in \mathcal{T}$ is estimated by a transition $\hat{t}_i \in \hat{\mathcal{T}}$. Let $\mathcal{S} \stackrel{\triangle}{=} \phi$ be a second estimate of the true transition path that assumes no transitions following $t_0 = 1$, i.e. $C(\mathcal{S}) = 0$.

Let the set $\bar{F}$ be the set of all N-tuples, for which $\left|\hat{t}_i - t_i\right| \leq 2.5k;\ \forall i,\ 1 \leq i \leq C$, i.e., all transitions are detected by $\hat{\mathcal{T}}$ near their true time point. Let F be the complimentary set of N-tuples. As figure B.1 depicts, if a transition t_i is detected at $\hat{t}_i$ more than $2.5k$ time points away from the true transition time, non-overlapping blocks are used to obtain $M(t_i)$ and $M\left(\hat{t}_i\right)$, while if $\left|\hat{t}_i - t_i\right| \leq 2.5k$, overlapping blocks may have been used for both metrics. For example: If true transition occurs in block EF, then block FG is used for its metric. Block GH, that is the first non-overlapping block, is used to detect transition in block HI. If transition is detected in HI, it is estimated at its midpoint, which is $2.5k + 1$ to $3.5k$ time points away from any point in EF, where the true transition occurs.

The average N-th order redundancy can be expressed as

$$R^N(\mathcal{D}) = \Pr(F) \cdot R^N(\mathcal{D} \mid F) + (1 - \Pr(F)) \cdot R^N(\mathcal{D} \mid \bar{F}), \tag{B.1}$$

where $R^N(\mathcal{D} \mid \mathcal{E})$ denotes the average N-th order redundancy given event $\mathcal{E}$ occurs. We can now treat each term separately to prove the theorem. We next present three propositions that upper bound different terms of (B.1), all assuming the conditions of the theorem. The proofs will be presented in Appendix C and Appendix D.

Proposition B.1 *The probability of F is upper bounded by*

$$\Pr(F) \leq C \cdot N \cdot 2^{-k\left[E(\Theta) - \frac{r-1}{k}\log(k+1)\right]}. \tag{B.2}$$

Proposition B.2

$$R_d\left(x^N\right) \leq 2.5C\frac{k\log N}{N} + O\left(\frac{Ck}{N}\right), \quad \forall x^N \in \bar{F}, \tag{B.3}$$

$$R\left(x^N\right) \leq 2.5AC\frac{\log^2 N}{N} + O\left(\frac{C\log N}{N}\right), \quad \forall x^N \in \bar{F}. \tag{B.4}$$

Proposition B.3

$$\Pr\left(\bar{F}\right)\cdot R_d^N\left(\mathcal{D} \mid \bar{F}\right) \leq 2.5\left[\log e + D\left(\Theta\right)\right]C\frac{k}{N} + o\left(\frac{Ck}{N}\right), \tag{B.5}$$

$$\Pr\left(\bar{F}\right)\cdot R^N\left(\mathcal{D} \mid \bar{F}\right) \leq K_1\frac{\log N}{N} + o\left(\frac{C\log N}{N}\right). \tag{B.6}$$

We will use the path $\mathcal{S}$ to estimate the transition paths of all N-tuples in F, for which $\hat{\mathcal{T}}$ is not a good estimate, and $\hat{\mathcal{T}}$ to estimate the paths of N-tuples in $\bar{F}$. Since the path $\mathcal{S}$ assumes the whole N-tuple is coded as a single stationary block, we can use eq. (16) with $m = N$ and $s = C+1$ to upper bound $R_d\left(x^N;\ \mathcal{D}\right)$,

$$R_d\left(x^N;\ \mathcal{D}\right) \leq \frac{m\log s}{N} = \log\left(C+1\right);\ \forall x^N \in F. \tag{B.7}$$

The path $\mathcal{S}$ assumes no transitions. Therefore, using Lemma A.1 we obtain TR of $O\left(\log N/N\right)$ and using eq. (14) we obtain PR of the same order. Thus

$$R\left(x^N;\ \mathcal{D}\right) \leq \log\left(C+1\right) + O\left(\frac{\log N}{N}\right);\ \forall x^N \in F. \tag{B.8}$$

Using this bound and Proposition B.1 and taking $k = A\log N$ and $\left(A\log N + 1\right) \leq \left(A+1\right)\log N$, we can upper bound the first term of eq. (B.1) by

$$\Pr\left(F\right)\cdot R^N\left(\mathcal{D} \mid F\right) \leq \left(A+1\right)^{r-1}C\frac{log^{r-1}N}{N^{AE(\Theta)-1}}\cdot\left[\log\left(C+1\right) + O\left(\frac{\log N}{N}\right)\right]. \tag{B.9}$$

Summing up eqs. (B.9) and (B.6) of Proposition B.3, we obtain an upper bound on the average redundancy for $D\left(\Theta\right) < \infty$. If $E\left(\Theta\right) > 2/A$ the term of (B.6) is dominant, thus obtaining the first region of the upper bound. If $1/A < E\left(\Theta\right) \leq 2/A$ the term of (B.9) is dominant, resulting in the third region of the upper bound. If $E\left(\Theta\right) \leq 1/A$ the upper bound of (B.9) is no longer useful. If $D\left(\Theta\right) = \infty$, we upper bound the probability of $\bar{F}$ by 1, and eq. (B.4) of Proposition B.2 results in the dominant term of the redundancy if $E\left(\Theta\right) > 2/A$, obtaining the second region of the bound. This concludes the proof of Theorem 3.

Appendix C – Proof of Proposition B.1

We begin the proof of Proposition B.1 with a few definitions. We then present and prove a lemma, upon which we base the proof of the proposition. Let $\hat{\mathcal{T}}$ be defined as in Appendix B, and let $t_i,\ 0 < i \leq C$ be the i-th true transition and s_i the block that contains t_i. The distribution before t_i is P_{i-1} and at t_i it becomes P_i. Now, let r be any block, s.t. $s_j + 1 < r < s_{j+1} - 1,\ 0 < j \leq C$, where for generalization we define $s_0 \stackrel{\triangle}{=} 0$ and $s_{\lfloor N/k \rfloor} \stackrel{\triangle}{=} \lfloor N/k \rfloor + 1$. Block r is defined s.t. blocks $r-1$, r and $r+1$ are entirely within the single true stationary segment j with distribution P_j.

Lemma C.1 *If $x^N \in F$ there exist s_i and r as defined above s.t. $M(s_i) \leq M(r)$, where both metrics are obtained from non-overlapping blocks.*

Proof: For transition in a block s_l, there are at most three states $s_l - 1$, s_l and $s_l + 1$ that may have metrics obtained by data of two different stationary segments. The DW scheme allows at most one of these three consecutive states to survive in the diagram and eliminates the other two. By definition of F, there exists $\hat{t}_i$, s.t. $\left|\hat{t}_i - t_i\right| > 2.5k$. Thus, there is a transition t_i, for which all these three states were eliminated. But since $\hat{C} = C$, this transition has an estimate. This estimate must be at some block r, s.t. r and its neighboring blocks are entirely within a stationary segment, because all surviving states with metrics obtained from two different segments are associated with other true transitions. Since s_i is eliminated and it is assumed the distance between transitions is larger than $O(k)$, it must be that s_i was eliminated because $M(s_i) \leq M(r)$. One of the blocks used to obtain $M(r)$ may still overlap a block used for a metric of some s_j, but $s_j \neq s_i$ since $\left|\hat{t}_i - t_i\right| > 2.5k$.

As a result of the lemma, we can bound $\Pr(F)$ by

$$\Pr(F) \leq \Pr\{\exists s_i, r \ : \ M(s_i) \leq M(r)\} \stackrel{\triangle}{=} \Pr(A). \tag{C.1}$$

We base the proof of the Proposition on this result. We select some fixed s_i and r, and upper bound the probability of event $A_{ir} \stackrel{\triangle}{=} \{M(s_i) \leq M(r)\}$. Then we use the union bound twice, once over all possible values of r and then on the values of i to upper bound the probability of the r.h.s. of inequality (C.1), which we denote as the probability of event A.

Observe some fixed s_i and r as defined above. For convenience, let us define $a \stackrel{\triangle}{=} s_i - 1$, $b \stackrel{\triangle}{=} s_i + 1$, $c \stackrel{\triangle}{=} r - 1$ and $d \stackrel{\triangle}{=} r + 1$. Figure C.1 illustrates this block partitioning. We define the empirical distribution P_α of block α as

$$P_\alpha(u) \stackrel{\triangle}{=} \frac{n_\alpha(u)}{k};\ \forall u \in \Sigma, \tag{C.2}$$

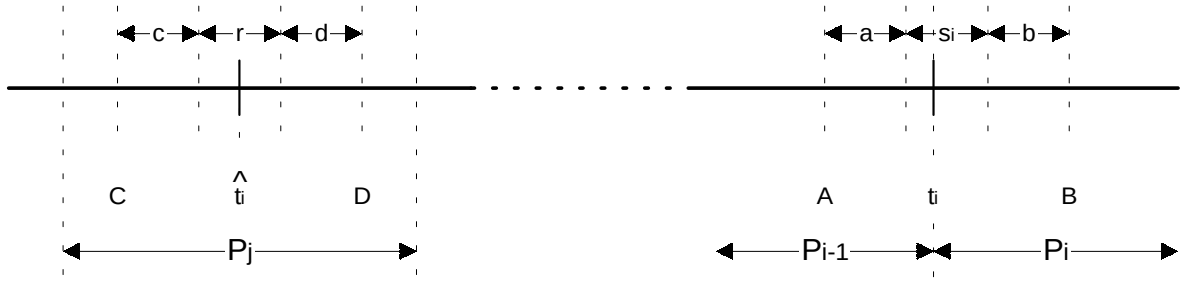

Figure C.1: typical occurrence of event F. The likelihood $M(r)$, obtained from blocks c and d both drawn by distribution P_j, is larger than $M(s_i)$, that represents the true transition t_i and is obtained from blocks a and b.

where $n_\alpha(u)$ is the number of occurrences of letter u in block α. Similarly, we define the empirical distribution of the concatenation of blocks α and β as

$$P_{\alpha\beta}(u) \triangleq \frac{n_\alpha(u) + n_\beta(u)}{2k};\ \forall u \in \Sigma. \tag{C.3}$$

The empirical per-letter entropies of a block and of a concatenation of two blocks are obtained by eqs. (36) and (37) respectively.

By definition of A_{ir} and $M(\cdot)$, we obtain that

$$A_{ir} = \left\{ x^N \ :\ H(c,d) - 0.5H(c) - 0.5H(d) \geq H(a,b) - 0.5H(a) - 0.5H(b) \right\}. \tag{C.4}$$

Rearranging terms, we can express A_{ir} by means of divergence

$$A_{ir} \triangleq \left\{ x_1^N \ :\ D(P_c||P_{cd}) + D(P_d||P_{cd}) \geq D(P_a||P_{ab}) + D(P_b||P_{ab}) \right\} \tag{C.5}$$

By typical sets analysis, (see [2] and [3]), and since blocks a,b,c and d are independent of each other, we can bound the probability of A_{ir} by

$$\Pr(A_{ir}) \leq 2^{-k\left[E(P_{i-1},P_i,P_j) - \frac{r-1}{k}\log(k+1)\right]}, \tag{C.6}$$

where $E(P_{i-1}, P_i, P_j)$ is obtained as

$$E(P_{i-1}, P_i, P_j) = \min_{A_{ir}} \{ D(P_a||P_{i-1}) + D(P_b||P_i) + D(P_c||P_j) + D(P_d||P_j) \}. \tag{C.7}$$

We will now define an event B_{ir}, for which the minimum of eq. (C.7) is easier to compute, s.t. $A_{ir} \subseteq B_{ir}$. Since A_{ir} is a subset of B_{ir}, the minimization over B_{ir} will lower bound the exponent defined by eq. (C.7). It is easy to show that

$$\begin{aligned} D(P_c||P_{cd}) + D(P_d||P_{cd}) &= D(P_c||P_j) + D(P_d||P_j) - 2D(P_{cd}||P_j) \\ &\leq D(P_c||P_j) + D(P_d||P_j). \end{aligned} \tag{C.8}$$

We define B_{ir} as

$$B_{ir} \stackrel{\triangle}{=} \left\{ x_1^N \ : \ D\left(P_c||P_j\right) + D\left(P_d||P_j\right) \geq D\left(P_a||P_{ab}\right) + D\left(P_b||P_{ab}\right) \right\}, \tag{C.9}$$

Since $A_{ir} \subseteq B_{ir}$,

$$\begin{aligned} E\left(P_{i-1}, P_i, P_j\right) &\geq \min_{B_{ir}} \left\{ D\left(P_a||P_{i-1}\right) + D\left(P_b||P_i\right) + D\left(P_c||P_j\right) + D\left(P_d||P_j\right) \right\} \\ &\geq \min_{\mathcal{C}} \left\{ D\left(P_a||P_{i-1}\right) + D\left(P_b||P_i\right) + D\left(P_a||P_{ab}\right) + D\left(P_b||P_{ab}\right) \right\} \\ &= \min_{\mathcal{C}} \left\{ \begin{array}{cc} \sum_{x\in\Sigma} P_a\left(x\right) \log \frac{2P_a^2(x)}{P_{i-1}(x)[P_a(x)+P_b(x)]} & + \\ \sum_{x\in\Sigma} P_b\left(x\right) \log \frac{2P_b^2(x)}{P_i(x)[P_a(x)+P_b(x)]} & \end{array} \right\}, \end{aligned} \tag{C.10}$$

where the constraint $\mathcal{C}$ is defined by

$$\sum_{x\in\Sigma} P_a\left(x\right) = \sum_{x\in\Sigma} P_b\left(x\right) = 1. \tag{C.11}$$

The second inequality is obtained by applying the constraint of B_{ir}, while the last equality is obtained by definition of divergence and by expressing $P_{ab}\left(x\right)$ as $0.5\left(P_a\left(x\right) + P_b\left(x\right)\right)$. The constrained minimization is performed using Lagrange multipliers. We define the functional $J\left(P_a, P_b\right)$ as

$$\begin{aligned} J\left(P_a, P_b\right) \ \stackrel{\triangle}{=} \ & \sum_{x\in\Sigma} P_a\left(x\right) \log \frac{2P_a^2\left(x\right)}{P_{i-1}\left(x\right)\left[P_a\left(x\right) + P_b\left(x\right)\right]} + \sum_{x\in\Sigma} P_b\left(x\right) \log \frac{2P_b^2\left(x\right)}{P_i\left(x\right)\left[P_a\left(x\right) + P_b\left(x\right)\right]} \\ & + \lambda_a \left(\sum_{x\in\Sigma} P_a\left(x\right) - 1 \right) + \lambda_b \left(\sum_{x\in\Sigma} P_b\left(x\right) - 1 \right). \end{aligned} \tag{C.12}$$

It is straightforward to show that the Hessian metrix of $J\left(P_a, P_b\right)$ w.r.t. $P_a\left(x\right)$ and $P_b\left(x\right)$ for $x \in \Sigma$ is positive definite. Hence the functional is convex and obtains the minimum where the first derivatives are zeros. By differentiation we obtain

$$E\left(P_{i-1}, P_i, P_j\right) \geq -2 \log \frac{\sum_{x\in\Sigma} \sqrt{P_{i-1}\left(x\right) P_i\left(x\right)} + 1}{2}. \tag{C.13}$$

Substituting the error exponent by its minimal value over all transitions, we can upper bound the probability of A_{ir} by

$$\Pr\left(A_{ir}\right) \leq 2^{-k\left[E(\Theta) - \frac{r-1}{k}\log(k+1)\right]}. \tag{C.14}$$

Using the union bound and accounting for all $\lfloor N/k - 3C \rfloor < N$ values of r first and then for all C values of i, we obtain the upper bound on $\Pr\left(A\right)$, and by using eq. (C.1) we apply this bound to $\Pr\left(F\right)$ and conclude the proof of Proposition B.1.

Appendix D – Proofs of Propositions B.2 and B.3

In this appendix we present the proofs of Propositions B.2 and B.3, that summarize the contribution of event $\bar{F}$ to the average redundancy. The difficulty in proving Proposition B.3 lies in the fact that the time instants of the estimates $\hat{t}_i$ vary for different N-tuples $x^N \in \bar{F}$, thus, we cannot assume anything about the distances $\left|\hat{t}_i - t_i\right|$ except that they are bounded by $2.5k$. We begin with analyzing the DR of an N-tuple $x^N \in \bar{F}$ by breaking it into $C+1$ terms, each represents the contribution of a single segment. We then break each such term into two different terms, which are analyzed separately. For each term we obtain a pointwise upper bound and an average one. Finally, we reconstruct the total DR by adding all the separate terms in a pointwise manner to prove Proposition B.2 and in the average to prove Proposition B.3. The second parts of both propositions are obtained by adding the PR and the TR to the DR. Throughout this section, we use the definition of $\hat{\mathcal{T}}$ presented in Appendix B.

We begin with some definitions. For $0 \leq i \leq C$, we define the following block lengths.

$$\begin{aligned} a_i &\stackrel{\triangle}{=} \max\left(t_i - \hat{t}_i, 0\right), \\ b_i &\stackrel{\triangle}{=} \max\left(\hat{t}_i - t_i, 0\right), \\ c_i &\stackrel{\triangle}{=} \min\left(t_{i+1}, \hat{t}_{i+1}\right) - \max\left(t_i, \hat{t}_i\right). \end{aligned} \tag{D.1}$$

We note that $a_0 = b_0 = 0$ and that for any i, either a_i or b_i must be zero. For generalization purposes we define $a_{C+1} = b_{C+1} \stackrel{\triangle}{=} 0$. We next define the vectors associated with each length.

$$\begin{aligned} x_a^i &\stackrel{\triangle}{=} \left(x_{\hat{t}_i}, x_{\hat{t}_i+1}, \ldots, x_{t_i-1}\right) \\ x_b^i &\stackrel{\triangle}{=} \left(x_{t_i}, x_{t_i+1}, \ldots, x_{\hat{t}_i-1}\right) \\ x_c^i &\stackrel{\triangle}{=} \left(x_{t_i+b_i}, x_{t_i+b_i+1}, \ldots, x_{t_{i+1}-a_{i+1}-1}\right). \end{aligned} \tag{D.2}$$

If the last index of a vector is smaller than the first, the vector will be the empty vector ϕ by definition. Therefore, for every i either x_a^i or x_b^i must be the empty vector. The probability of the empty vector with any distribution will be defined as 1. The nonempty vectors obtained from the $C+1$ sets of the above three vectors are a complete parsing of the N-tuple into disjoint strings. Hence, we can express the DR as the sum of the contributions of all these vectors, where the contribution of ϕ is zero. Finally, the empirical distribution of vector $x_\alpha^i \neq \phi$ is defined as

$$P_\alpha^i(u) \stackrel{\triangle}{=} \frac{n_\alpha^i(u)}{\alpha_i};\ \forall u \in \Sigma, \tag{D.3}$$

where $n_\alpha^i(u)$ is the number of occurrences of u in x_α^i. Figure D.1 demonstrates the definitions presented above. Each sting x_α^i is noted by its length α_i.

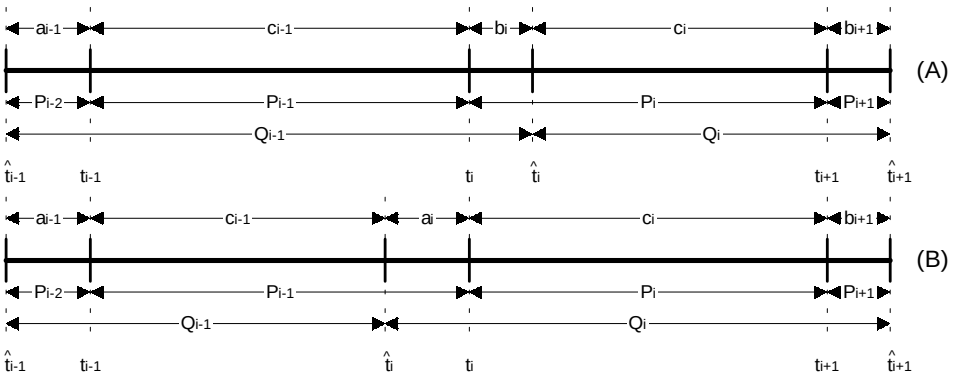

Figure D.1: typical occurrence of event $\bar{F}$ and its effect on the estimated distributions. The dark line represents the time axis, that is partitioned into strings, whose length is noted above the dark line. The true distributions of the segments are noted by P and the distributions assigned to the hypothesized segments by the selection of $\hat{\mathcal{T}}$ are noted by Q. In diagram (A) $\hat{t}_i > t_i$ and in diagram (B) $\hat{t}_i < t_i$. In both diagrams $\hat{t}_{i-1} < t_{i-1}$ and $\hat{t}_{i+1} > t_{i+1}$.

We can now use the above definitions to define the distributions Q_i that the path $\hat{\mathcal{T}}$ assigns to each of its hypothesized segments. For convenience, we define

$$A_i \stackrel{\triangle}{=} a_i + c_i + b_{i+1}, \tag{D.4}$$

and then

$$Q_i(u) \stackrel{\triangle}{=} \frac{1}{A_i}\left[a_i P_{i-1}(u) + c_i P_i(u) + b_{i+1} P_{i+1}(u)\right] \;;\; 0 \le i \le C,\ \forall u \in \Sigma, \tag{D.5}$$

where P_i is the true distribution of segment i. The distributions P_{-1} and P_{C+1} need not be defined, since they will always be multiplied by zero. Figure D.1 illustrates the true and the hypothesized distributions around transition t_i for both cases $\hat{t}_i > t_i$ and $\hat{t}_i < t_i$. In both diagrams we assume $a_{i-1} > 0$ and $b_{i+1} > 0$.

We can now represent the DR of $x^N \in \bar{F}$ as the sum of $C+1$ terms (one for each segment), each consists of two terms, the first is the contribution of vectors a, b and the second of vector c.

$$\begin{aligned} R_d\left(x^N; \mathcal{D}\right) &\stackrel{\triangle}{=} \frac{1}{N}\log\frac{P\left(x^N \mid \Theta, \mathcal{T}\right)}{Q\left(x^N \mid \hat{\Theta}, \hat{\mathcal{T}}\right)} \\ &= \frac{1}{N}\sum_{i=0}^{C}\log\frac{P_{i-1}\left(x_a^i\right) P_i\left(x_b^i\right) P_i\left(x_c^i\right)}{Q_i\left(x_a^i\right) Q_{i-1}\left(x_b^i\right) Q_i\left(x_c^i\right)} \\ &= \frac{1}{N}\sum_{i=0}^{C}\log\frac{P_{i-1}\left(x_a^i\right) P_i\left(x_b^i\right)}{Q_i\left(x_a^i\right) Q_{i-1}\left(x_b^i\right)} + \frac{1}{N}\sum_{i=0}^{C}\log\frac{P_i\left(x_c^i\right)}{Q_i\left(x_c^i\right)} \\ &\stackrel{\triangle}{=} \sum_{i=0}^{C} R_i + \sum_{i=0}^{C} r_i = \sum_{i=1}^{C} R_i + \sum_{i=0}^{C} r_i. \end{aligned} \tag{D.6}$$

The equality in the second line of (D.6) is obtained by definition of strings x_a^i, x_b^i and x_c^i in (D.2). The string x_a^i is drawn by P_{i-1} but is assumed to be drawn by Q_i. The other two strings x_b^i and

x_c^i are drawn by P_i, but are assumed to be drawn by Q_{i-1} and Q_i respectively. The last equality is obtained by noticing that $x_a^0 = x_b^0 = \phi$. We summarize the pointwise and average bounds of r_i and R_i in the following lemma.

Lemma D.1

$$r_i \leq \frac{a_i + b_{i+1}}{N} \log e; \quad 0 \leq i \leq C, \tag{D.7}$$

$$R_i \leq \frac{(a_i + b_i) \log N}{N}; \quad 1 \leq i \leq C, \tag{D.8}$$

$$\Pr\left(\bar{F}\right) E\left[R_i \mid \bar{F}\right] \leq \frac{2.5k}{N}\left[D\left(P_{i-1}||P_i\right) + D\left(P_i||P_{i-1}\right)\right] + o\left(\frac{k}{N}\right); \quad 1 \leq i \leq C. \tag{D.9}$$

The proof of the lemma is presented at the end of this section. Using eq. (D.7) and noting that $a_0 = b_{C+1} = 0$, we obtain

$$\sum_{i=0}^{C} r_i \leq \sum_{i=1}^{C} \frac{a_i + b_i}{N} \log e \leq 2.5 \left(\log e\right) C \frac{k}{N}, \tag{D.10}$$

where the last inequality is obtained by the fact that for each i either a_i or b_i must be zero and the definition of $\bar{F}$ that ensures that either must be bounded by $2.5k$. Similarly, we can show that

$$\sum_{i=1}^{C} R_i \leq 2.5C \frac{k \log N}{N}. \tag{D.11}$$

Adding both bounds of the last two equations, we conclude the proof of the first equation of Proposition B.2. The proof of Proposition B.2 is concluded by simply adding the bounds for the PR and TR in eq. (14) and (A.1) respectively with $\hat{C} = C$ and taking $k = A \log N$, noticing the DR is the dominant term. Proposition B.3 is proved similarly to Proposition B.2 with one difference. Instead of taking the bound on R_i of eq. (D.8), we take the upper bound of (D.9) for the average R_i over all transitions to obtain $D\left(\Theta\right)$. To conclude this section, we present the proof of Lemma D.1.

Proof of Lemma D.1: Eq. (D.7) and (D.8) are proved by straightforward manipulations. Before proving eq. (D.9) we present and prove another lemma. We can upper bound r_i in the following manner.

$$r_i \triangleq \frac{1}{N} \log \frac{P_i\left(x_c^i\right)}{Q_i\left(x_c^i\right)} \tag{D.12}$$

$$= \frac{c_i}{N} \sum_{x \in \Sigma} P_c^i\left(x\right) \log \frac{P_i\left(x\right)}{Q_i\left(x\right)} \tag{D.13}$$

$$\leq \frac{c_i}{N} \sum_{x \in \Sigma} P_c^i\left(x\right) \log \frac{P_i\left(x\right)}{\frac{c_i}{A_i} P_i\left(x\right)} \tag{D.14}$$

$$= \frac{c_i}{N} \log \frac{A_i}{c_i} \tag{D.15}$$

$$\leq \frac{1}{N} \lim_{c_i \to \infty} c_i \log \frac{A_i}{c_i} = \frac{a_i + b_{i+1}}{N} \log e. \tag{D.16}$$

Eq. (D.13) is obtained by representing the probability of vector x_c^i as the sum of the probabilities of its components. Using the definition of Q_i in (D.5) we obtain inequality (D.14). We note that the function on the l.h.s. of inequality (D.16) is an increasing function of c_i in order to upper bound it by its value for $c_i \to \infty$. Finally, we use L'hospital rule to obtain this value, concluding the proof of eq. (D.7). We perform similar analysis to show that

$$\begin{aligned} R_i &\leq \frac{a_i}{N} \log \frac{A_i}{a_i} + \frac{b_i}{N} \log \frac{A_i}{b_i} \\ &\leq \frac{a_i + b_i}{N} \log N. \end{aligned} \tag{D.17}$$

The second inequality is obtained by taking $A_i \to N$ and discarding the denominators of the terms inside the logarithms, thus concluding the proof of eq. (D.8).

Lemma D.2

$$\begin{aligned} \frac{1}{N} \log \frac{P_{i-1}\left(x_a^i\right)}{Q_i\left(x_a^i\right)} &\leq \frac{1}{N} \log \frac{P_{i-1}\left(x_a^i\right)}{P_i\left(x_a^i\right)} + o\left(\frac{k}{N}\right), \\ \frac{1}{N} \log \frac{P_i\left(x_b^i\right)}{Q_{i-1}\left(x_b^i\right)} &\leq \frac{1}{N} \log \frac{P_i\left(x_b^i\right)}{P_{i-1}\left(x_b^i\right)} + o\left(\frac{k}{N}\right). \end{aligned} \tag{D.18}$$

Proof of Lemma D.2: We prove the first inequality, but the same analysis can be performed to prove the second. We first upper bound the expression $\log\left(P_{i-1}(x)/Q_i(x)\right)$ for $x \in \Sigma$ where $P_{i+1}(x) > 0$.

$$\begin{aligned} \log \frac{P_{i-1}(x)}{Q_i(x)} &= \log \frac{P_{i-1}(x)}{\frac{1}{A_i}\left[a_i P_{i-1}(x) + c_i P_i(x) + b_{i+1} P_{i+1}(x)\right]} \\ &\leq \frac{c_i}{A_i} \log \frac{P_{i-1}(x)}{P_i(x)} + \frac{b_{i+1}}{A_i} \log \frac{P_{i-1}(x)}{P_{i+1}(x)} \\ &\leq \log \frac{P_{i-1}(x)}{P_i(x)} + o(1). \end{aligned} \tag{D.19}$$

The first inequality is obtained by Jensen's inequality and the second by the assumption that transitions are larger than $O(k)$ apart, thus $O(A_i) > O(b_{i+1})$. We now show that we can obtain the same bound when $P_{i+1}(x) = 0$. We define $B_i \triangleq a_i + c_i$.

$$\begin{aligned} \log \frac{P_{i-1}(x)}{Q_i(x)} &= \log \frac{\frac{A_i}{B_i} P_{i-1}(x)}{\frac{a_i}{B_i} P_{i-1}(x) + \frac{c_i}{B_i} P_i(x)} \\ &\leq \log \frac{A_i}{B_i} + \frac{c_i}{B_i} \log \frac{P_{i-1}(x)}{P_i(x)} \leq \log \frac{P_{i-1}(x)}{P_i(x)} + o(1). \end{aligned} \tag{D.20}$$

The first inequality is obtained by Jensen's inequality, and the second by the fact that $O\left(B_i\right) > O\left(b_{i+1}\right)$, since transitions are larger than $O\left(k\right)$ apart. We obtain the inequality by the well known fact that if $x \to 0$, $\log\left(1+x\right) = o\left(x\right)$. Summing up for all terms of x_a^i we obtain that

$$\begin{aligned}
\frac{1}{N}\log\frac{P_{i-1}\left(x_a^i\right)}{Q_i\left(x_a^i\right)} &= \frac{a_i}{N}\sum_{x\in\Sigma}P_a^i\left(x\right)\log\frac{P_{i-1}\left(x\right)}{Q_i\left(x\right)} \\
&\leq \frac{a_i}{N}\sum_{x\in\Sigma}P_a^i\left(x\right)\left[\log\frac{P_{i-1}\left(x\right)}{P_i\left(x\right)}+o\left(1\right)\right] \\
&= \frac{1}{N}\log\frac{P_{i-1}\left(x_a^i\right)}{P_i\left(x_a^i\right)}+o\left(\frac{k}{N}\right),
\end{aligned} \tag{D.21}$$

which concludes the proof of Lemma D.2.

To conclude the proof of Lemma D.1, we first extend the definition of a_i and b_i s.t.

$$a_i = b_i = 0;\ 1 \leq i \leq C,\quad \forall x^N \in F. \tag{D.22}$$

The respective strings x_a^i and x_b^i are defined as the null strings. We define $\alpha \stackrel{\triangle}{=} 2.5k$, and for all i, s.t. $1 \leq i \leq C$, we make the following definitions.

$$\begin{aligned}
u_i &\stackrel{\triangle}{=} \alpha - a_i, \\
v_i &\stackrel{\triangle}{=} \alpha - b_i, \\
x_u^i &\stackrel{\triangle}{=} \left(x_{t_i-\alpha}, x_{t_i-\alpha+1}, \ldots, x_{t_i-a_i-1}\right), \\
x_v^i &\stackrel{\triangle}{=} \left(x_{t_i+b_i}, x_{t_i+b_i+1}, \ldots, x_{t_i+\alpha-1}\right).
\end{aligned} \tag{D.23}$$

The idea is to create the concatenated strings x_{ua}^i and x_{bv}^i, s.t. the first will contain all the α letters right before the i-th true transition and the second the α letters right after the transition, for all $x^N \in \Sigma^N$. By using this notation, we can generalize the analysis for R_i and average over x^N. We conclude by showing the proof of eq. (D.9), for convenience, we omit the superscript i from all vectors.

$$\Pr\left(\bar{F}\right)E\left[R_i|\bar{F}\right] = \frac{1}{N}\sum_{x^N\in\bar{F}}\Pr\left(x^N\right)\log\frac{P_{i-1}\left(x_a\right)P_i\left(x_b\right)}{Q_i\left(x_a\right)Q_{i-1}\left(x_b\right)} \tag{D.24}$$

$$\leq \frac{1}{N}\sum_{x^N\in\bar{F}}\Pr\left(x^N\right)\log\frac{P_{i-1}\left(x_a\right)P_i\left(x_b\right)}{P_i\left(x_a\right)P_{i-1}\left(x_b\right)}+o\left(\frac{k}{N}\right) \tag{D.25}$$

$$\leq \frac{1}{N}\sum_{x^N\in\bar{F}}\Pr\left(x^N\right)\left[\left|\log\frac{P_{i-1}\left(x_a\right)}{P_i\left(x_a\right)}\right|+\left|\log\frac{P_i\left(x_b\right)}{P_{i-1}\left(x_b\right)}\right|\right]+o\left(\frac{k}{N}\right) \tag{D.26}$$

$$\begin{aligned}
\leq \frac{1}{N}\sum_{x^N\in\Sigma^N}\Pr\left(x^N\right)\Bigg[&\left|\log\frac{P_{i-1}\left(x_a\right)}{P_i\left(x_a\right)}\right|+\left|\log\frac{P_i\left(x_b\right)}{P_{i-1}\left(x_b\right)}\right| \\
&+\left|\log\frac{P_{i-1}\left(x_u\right)}{P_i\left(x_u\right)}\right|+\left|\log\frac{P_i\left(x_v\right)}{P_{i-1}\left(x_v\right)}\right|\Bigg]+o\left(\frac{k}{N}\right)
\end{aligned} \tag{D.27}$$

$$
\begin{aligned}
&\leq \frac{1}{N}\sum_{x^N\in\Sigma^N}\Pr\left(x^N\right)\left[\log\frac{P_{i-1}\left(x_a\right)}{P_i\left(x_a\right)}+\frac{2\log e}{e}\frac{P_i\left(x_a\right)}{P_{i-1}\left(x_a\right)}\right. && \text{(D.28)}\\
&\quad+\log\frac{P_i\left(x_b\right)}{P_{i-1}\left(x_b\right)}+\frac{2\log e}{e}\frac{P_{i-1}\left(x_b\right)}{P_i\left(x_b\right)}\\
&\quad+\log\frac{P_{i-1}\left(x_u\right)}{P_i\left(x_u\right)}+\frac{2\log e}{e}\frac{P_i\left(x_u\right)}{P_{i-1}\left(x_u\right)}\\
&\quad\left.+\log\frac{P_i\left(x_v\right)}{P_{i-1}\left(x_v\right)}+\frac{2\log e}{e}\frac{P_{i-1}\left(x_v\right)}{P_i\left(x_v\right)}\right]+o\left(\frac{k}{N}\right)\\
&= \frac{1}{N}\sum_{x_{ua}\in\Sigma^\alpha}P_{i-1}\left(x_{ua}\right)\log\frac{P_{i-1}\left(x_{ua}\right)}{P_i\left(x_{ua}\right)}+ && \text{(D.29)}\\
&\quad\frac{1}{N}\sum_{x_{bv}\in\Sigma^\alpha}P_i\left(x_{bv}\right)\log\frac{P_i\left(x_{bv}\right)}{P_{i-1}\left(x_{bv}\right)}+o\left(\frac{k}{N}\right)\\
&= \frac{2.5k}{N}\left[D\left(P_{i-1}||P_i\right)+D\left(P_i||P_{i-1}\right)\right]+o\left(\frac{k}{N}\right). && \text{(D.30)}
\end{aligned}
$$

Inequality (D.25) is obtained by Lemma D.2. We then upper bound the two logarithm terms by their absolute value to obtain eq. (D.26), and add non-negative terms for (D.27). The bound $|\log x|\leq 2\frac{\log e}{e}\frac{1}{x}+\log x$ (see [5]) is then used to obtain (D.28). Finally, we discard all data letters independent of x_{ua} and x_{bv} and realize that the distributions of x_{ua} and x_{bv} are P_{i-1} and P_i respectively (eq. (D.29)), and hence we obtain the sum of divergences in (D.30) multiplied by the length of the strings $2.5k$, concluding the proof of (D.9).

Appendix E – Proof of Theorem 4

To prove Theorem 4 we first need to upper bound the number of blocks B. We then use this bound to upper bound both the PR and the DR. By definition of the scheme

$$
\begin{aligned}
N &\geq 1+\sum_{b=2}^{B-1}\lfloor kb\log b\rfloor\geq\sum_{b=1}^{B-1}\left[kb\log b-1\right] && \text{(E.1)}\\
&\geq k\int_0^{B-1}x\log x dx-B=\frac{k\log e}{2}\left[x^2\left(\ln x-0.5\right)\right]_0^{B-1}-B\\
&= \frac{k}{2}\left[\left(B-1\right)^2\log\frac{B-1}{\sqrt{e}}\right]-B=\frac{k}{2}B^2\log B-O\left(B^2\right).
\end{aligned}
$$

Therefore,

$$
N+O\left(B^2\right)\geq\frac{k}{2}B^2\log B. \tag{E.2}
$$

To satisfy the last equation, we must have

$$
B\leq\sqrt{\frac{4}{k}}\sqrt{\frac{N}{\log N}}+o\left(\sqrt{\frac{N}{\log N}}\right). \tag{E.3}
$$

As a result of the discussion before eq. (14), each block of length m_b, coded by the KT estimate, produces $\frac{r-1}{2}\log m_b + O(1)$ extra PR bits. Hence, we can upper bound the PR by

$$\begin{aligned} R_p\left(x^N;\mathcal{P}\right) &\leq \frac{r-1}{2N}\sum_{b=1}^{B}\left[\log\lfloor kb\log b\rfloor + O(1)\right] && \text{(E.4)}\\ &\leq \frac{r-1}{2N}\left[\sum_{b=1}^{B}\log\left(kb\log b\right) + O(B)\right] && \text{(E.5)}\\ &= \frac{r-1}{2N}\left[B\log k + \log(B!) + \sum_{b=1}^{B}\log\log b + O(B)\right] && \text{(E.6)}\\ &\leq \frac{r-1}{2N}\left[B\log B + O(B\log\log B)\right] = \frac{r-1}{2\sqrt{k}}\sqrt{\frac{\log N}{N}} + o\left(\sqrt{\frac{\log N}{N}}\right). && \text{(E.7)} \end{aligned}$$

Eq. (E.6) is obtained by opening the expression inside the logarithm, and inequality (E.7) by the fact that $B! \leq B^B$.

It is obvious from (16) that the largest DR is obtained for PSMS's with true transitions at the midpoints of the last C blocks. For these PSMS's we can upper bound the DR by

$$R_d\left(x^N;\mathcal{P}\right) \leq C\frac{m_B}{N} \leq Ck\frac{B\log B}{N} = C\sqrt{k}\sqrt{\frac{\log N}{N}} + o\left(\sqrt{\frac{\log N}{N}}\right). \qquad \text{(E.8)}$$

We conclude the proof of Theorem 4 by summing up the last two upper bounds.

Acknowledgment

The authors gratefully acknowledge Professor J. Ziv, who served as the first author's M.Sc. co-adviser, for his helpful suggestions.

References

[1] R. E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, 1991.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[3] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.*, Academic Press, New York, 1981.

[4] R. G. Gallager, "Variations on a theme by Huffman," *IEEE Trans. Inform. Theory*, Vol. IT-24, pp. 668-674, Nov. 1978.

[5] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York, 1990.

[6] M. Gutman, "Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics," *IEEE Trans. Inform. Theory*, Vol. 35, No. 2, pp. 401-408, March 1989.

[7] P. G. Howard and J. S. Vitter, "Arithmetic coding for data compression," *Proc. of the IEEE*, Vol. 82, No. 6, pp. 857-865, June 1994.

[8] F. Jelinek, *Probabilistic Information Theory*, New York: McGraw-Hill, 1968, pp. 476-489.

[9] D. E. Knuth, "Dynamic Huffman coding," *J. Algorithms*, Vol. 6, pp. 163-180, June 1985.

[10] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, Vol. IT-27, pp. 199-207, March 1981.

[11] N. Merhav, "On the minimum description length principle for sources with piecewise constant parameters," *IEEE Trans. Inform. Theory*, Vol. 39, No. 6, pp. 1962-1967, November 1993.

[12] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, Vol. 41, No. 3, pp. 714-722, May 1995.

[13] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, Vol. IT-30, No. 4, pp. 629-636, July 1984.

[14] Y. M. Shtarkov, T. J. Tjalkens and F. M. J. Willems, "Multi-alphabet universal coding of memoryless sources," *Problems of Information Transmission*, Vol. 31, No. 2, pp 20-35, April-June, 1995.

[15] J. S. Vitter, "Design and analysis of dynamic Huffman codes," *J. Ass. Comput. Mach.*, Vol. 34, pp. 825-845, Oct. 1987.

[16] F. M. J. Willems, "Coding for binary piecewise memoryless sources," *Proc. of Japan Benelux Workshop*, 1994.

[17] F. M. J. Willems, "Coding for a binary Independent Piecewise-Identically-Distributed source," *IEEE Trans. Inform. Theory*, Vol. 42, No. 6, pp. 2210-2217, November 1996.

[18] F. M. J. Willems and F. Casadei, "Weighted coding methods for binary piecewise memoryless sources," *Proceedings of 1995 IEEE International Symposium on Information Theory*, Canada September 17-22, 1995.

[19] F. M. J. Willems and M. Krom, "Live-and-die coding for binary piecewise i.i.d. sources," *Proceedings of 1997 IEEE International Symposium on Information Theory*, pp. 68, Ulm, Germany, June 29 - July 4, 1997.

[20] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, "The Context-Tree weighting method: basic properties," *IEEE Trans. Inform. Theory*, Vol. 41, No. 3, pp. 653-664, May 1995.

[21] A. D. Wyner and J. Ziv, "The sliding-window Lempel-Ziv algorithm is asymptotically optimal," *Proc. of the IEEE*, Vol. 82, No. 6, pp. 872-877 June 1994.

[22] H. Yokoo, "An improvement of dynamic Huffman coding with a simple repetition finder," *IEEE Trans. Commun.*, Vol. 39, pp. 8-10, Jan. 1991.

[23] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, Vol. 34, No. 2, pp. 278-286, March 1988.

[24] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, Vol. IT-24, pp. 337-343, May 1977.

[25] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, Vol IT-24, pp. 530-536, Sept. 1978.