

# Delay and Redundancy in Lossless Source Coding

Ofer Shayevitz, Eado Meron, Meir Feder and Ram Zamir

**Abstract**—The penalty incurred by imposing a finite delay constraint in lossless source coding of a memoryless source is investigated. It is well known that for the so-called block-to-variable and variable-to-variable codes, the redundancy decays at best polynomially with the delay, where in this case the delay is identified with the source block length or maximal source phrase length, respectively. In stark contrast, it is shown that for sequential codes (e.g., a delay-limited arithmetic code) the redundancy can be made to decay exponentially with the delay constraint. The corresponding redundancy–delay exponent is shown to be at least as good as the Rényi entropy of order 2 of the source, but (for almost all sources) not better than a quantity depending on the minimal source symbol probability and the alphabet size.

## I. INTRODUCTION

It is well known that any memoryless source can be asymptotically losslessly compressed to its entropy [1]. However, in the presence of resource constraints, a rate penalty, referred to as *redundancy*, is unavoidable. In this work we focus on the redundancy in the encoding of a memoryless source incurred by the imposition of a *strict end-to-end delay constraint*  $d$  measured in source clocks, i.e., under the requirement that the  $n$ -th encoded symbol must always be perfectly reproduced at the decoder by time  $n + d$ .

In the lossless source coding literature, three classes of codes in which delay is a design parameter are traditionally studied: 1) The Block-to-Variable (BV) class (e.g. Huffman code [2]), where a source sequence is partitioned into equal-length blocks and each block is mapped to a unique variable length codeword from a prefix-free set, 2) The Variable-to-Block (VB) class (e.g. Tunstall code [3], [4]), where the source sequence is parsed into phrases according to a complete code-tree, and each phrase is mapped to a unique fixed length codeword, and 3) The Variable-to-Variable (VV) class (e.g., Khodak codes), where the source sequence is similarly parsed but each phrase is mapped to a unique variable length codeword from a prefix-free set. In the sequel, we collectively refer to the three classes above as *the classical framework*. In the BV class, a delay constraint is interpreted as a block length constraint, and the redundancy is known to decay at best polynomially with the delay [5], [6]. In the VB/VV class (where the delay is a random variable depending on the source sequence) the delay constraint is translated into a maximal phrase length constraint, and the redundancy again decays at

best polynomially with the delay, though sometimes faster than in the BV case [4], [7], [8]<sup>1</sup>.

In a delay constrained setting, the classical framework admits two (related) limitations. First, even within that framework, there is an apparent disparity between delay and block/phrase length. The reason block/phrase lengths are identified with delay in the first place is since concatenating code-words allows the source reproduction at block/phrase length intervals. However, the delay can sometimes be significantly shorter, for essentially the same reason: Consider a BV code of block length  $n = kd$  obtained by concatenating  $k$  BV codes of block length  $d$ . Clearly, the decoder can reproduce symbols with a delay  $d$ , rather than the possibly much larger delay  $n$ . Waiting until the end of the block would mean the encoder is “holding back” bits it is already certain of, clearly an undesirable trait in a delay constrained setting. Of course, the redundancy associated with such an encoder in the limit of  $k \rightarrow \infty$  still decays polynomially as a function of  $d$ , which brings us to the second limitation. In the memoryless classical framework, the encoder never looks beyond the end of the current block/phrase, in the sense that the source’s prefix has no effect on the output of the encoder beyond that point<sup>2</sup>. The encoder is therefore being “reset” roughly every  $d$  symbols. Loosely speaking, the penalty incurred by forcing these regularly recurring reset points, is the source of the polynomial delay of the redundancy.

With these observations in mind, we recall a lossless coding technique of a different flavor that does not suffer from the above shortcomings. In *arithmetic coding* [10], [11], [12], [13], a source subsequence is sequentially mapped into nested subintervals of the unit interval, with length equal to the sequence probability, and the common most significant bits of the current subinterval are emitted. This way, the encoder never holds back any bits it is already certain of, by definition. Moreover, whereas BV/VB/VV encoders never look beyond the end of the current block/phrase, an arithmetic encoder constantly looks into the (possibly infinite) future. Unfortunately, this comes at a cost of an unbounded delay (though a bounded expected delay, see [14], [15], [16]). Nevertheless, the notion of arithmetic coding does point us in the right direction. In a delay constrained framework, an encoder should *by definition* be sequential, emitting all the bits it can at any given instance. Moreover, a good delay constrained encoder should always strive to look  $d$  steps ahead, avoiding “reset” points as much as possible. As we shall see, these properties are nicely captured within an interval mapping type framework.

In this paper, we introduce a general framework for lossless delay constrained coding of a memoryless source, and study

The authors are with the Department of EE-Systems, Tel Aviv University, Tel Aviv, Israel {email: ofersha@eng.tau.ac.il, meroneado@gmail.com, meir@eng.tau.ac.il, zamir@eng.tau.ac.il}. This paper was presented in part at ISIT 2006, DCC 2007 and DCC 2008. The work of O. Shayevitz was partially supported by the Adams Fellowship and the ITA fellowship. The work of R. Zamir was partially supported by the Israel Academy of Science, ISF grant number 870/11.

<sup>1</sup>These results hold even in the weaker case of an expected delay constraint.

<sup>2</sup>This assertion does not hold for sources with memory, where dependencies between phrases can be beneficial [9].

the fundamental tradeoff between delay and redundancy. We show that, in stark contrast to the polynomial decay within the classical framework, the redundancy  $\mathfrak{R}(P, d)$  associated with a memoryless source  $P$  over a finite alphabet  $\mathcal{X}$ , can be made to decay *exponentially* with the delay  $d$ . Specifically, we show that any encoder obeying a delay constraint  $d$  satisfies<sup>3</sup>

$$\left(\frac{p_{\min}}{|\mathcal{X}|}\right)^{8d} \lesssim \mathfrak{R}(P, d) \lesssim p_{\max}^d$$

where  $p_{\min}, p_{\max}$  are the minimal and maximal source symbol probabilities, the upper bound holds for all sources, and the lower bound holds for almost all sources<sup>4</sup>. We then tighten the upper bound and obtain

$$\mathfrak{R}(P, d) \lesssim 2^{-dH_2(P)}$$

where  $H_2(P)$  is the Rényi entropy of order 2 of the source. For our upper bound, we introduce a construction based on mismatched arithmetic coding in conjunction with a fictitious symbol insertion mechanism. For our lower bound, we provide a “generalized interval mapping” representation for delay constrained encoders.

*Related work.* Whereas in this paper we consider the impact of an end-to-end delay constraint measured in source clocks, other works have considered complementary questions where delay is measured in encoded bits. In [15], [19] the authors describe a variable-length lossless source coding system based on finite precision arithmetic coding, that falls outside the classical framework and is of a similar flavor to the codes considered herein; Specifically, they show [19, Appendix II] that the associated redundancy decays exponentially with the maximal number of encoded bits the decoder can hold in its queue. A similar observation can be deduced from the discussion in [20]. While employing a different measure of delay, it appears plausible (but remains unverified) that these constructions could also be employed to derive an exponential upper bound on the redundancy as a function of the delay in source clocks. None of these prior works provided a lower bound for the redundancy. In [21], the author considers a setting where the channel connecting the encoder and the decoder can transmit a fixed number of bits per second, and has a finite length queue at its input. He shows that the probability of queue overflow for BV codes can be made to decay exponentially with the size of the queue, and describes the tradeoff between the exponent and the minimal achievable compression rate.

*Organization.* Our framework is introduced in Section II, and some basic lemmas are derived. In Section III, the delay profile of mismatched arithmetic coding is analyzed. This analysis is then applied in Section IV where a lower bound on the redundancy–delay exponent is derived. In Section V, a corresponding upper bound on the redundancy–delay exponent for almost all sources is presented. Some final remarks are given in Section VI.

<sup>3</sup>By  $a_d \lesssim b_d$  we mean  $\liminf_{d \rightarrow \infty} \frac{1}{d} \log \frac{b_d}{a_d} > 0$ .

<sup>4</sup>Recall that the reason for jointly coding over multiple source symbols, and consequently incurring delay, is to make the rounding error of the log-probabilities negligible. This is unnecessary for dyadic sources, where symbol probabilities are all integer powers of 2. Hence, a lower bound cannot hold for all sources, as dyadic sources can attain zero redundancy with zero delay.

## II. PRELIMINARIES

### A. Notations

We write  $s \preceq t$  to indicate that a string  $s$  is a prefix of a string  $t$ , and  $s \prec t$  to indicate that  $s \preceq t$  and  $s \neq t$ . A set of finite strings  $S$  is said to be *prefix-free* if no pair of strings  $s, t \in S$  satisfies  $s \prec t$ . The *longest common prefix* of  $S$  is the string  $t$  of maximal length satisfying  $t \preceq s$  for all  $s \in S$ . The Lebesgue measure of a set  $A \subseteq \mathbb{R}$  is denoted by  $|A|$ . The *fractional part* of a number  $a \in \mathbb{R}$  is denoted by  $\langle a \rangle \stackrel{\text{def}}{=} a - \lfloor a \rfloor$ . The *difference modulo-1*  $\langle A - B \rangle$  between two sets  $A, B \subseteq \mathbb{R}$  is the set of all numbers  $\langle a - b \rangle$  where  $a \in A, b \in B$ . For any function  $f : \mathbb{R} \mapsto \mathbb{R}$  and any set  $A \subseteq \mathbb{R}$ , we write  $f(A)$  for the image of  $A$  under  $f$ . All logarithms are taken to the base of 2. A *total order* of a finite set is called simply an *order*.

The following lemma is easily verified.

**Lemma 1.** *Let  $A, B \subseteq \mathbb{R}$  be any two sets. Then*

- (i) *If  $b \in B$  and  $\langle c \rangle \notin \langle A - B \rangle$ , then  $b + c \notin A$ .*
- (ii) *If  $b \in B$  and  $\langle \log c \rangle \notin \langle \log A - \log B \rangle$ , then  $bc \notin A$ .*

### B. Sources

Let  $\mathcal{X}$  be a finite alphabet of source symbols. The set of all length- $n$  strings of symbols from  $\mathcal{X}$  is denoted  $\mathcal{X}^n$ , the set of all finite length strings is denoted  $\mathcal{X}^*$ , and the set of all infinite length strings is denoted  $\mathcal{X}^\infty$ . We sometimes use the notations  $x^n \stackrel{\text{def}}{=} x_1 x_2 \dots x_n$  and  $x_m^n \stackrel{\text{def}}{=} x_m x_{m+1} \dots x_n$  for finite source strings, where the convention is that  $x_m^n = \emptyset$  when  $m > n$ . A *discrete memoryless source (DMS)*  $P$  is defined by a *probability mass function (p.m.f.)*  $\{P(x) : x \in \mathcal{X}\}$  which naturally induces a product measure over  $\mathcal{X}^*$ , via  $P(st) = P(s)P(t)$  for all  $s, t \in \mathcal{X}^*$ , where  $st$  is the concatenation of  $s$  and  $t$ . Specifically, we denote by  $P^n$  the p.m.f. obtained by restricting  $P$  to  $\mathcal{X}^n$ . An infinite random source string emitted by the source  $P$  will be denoted by  $X^\infty$ . The minimal and maximal symbol probabilities under  $P$  are denoted  $p_{\min}$  and  $p_{\max}$  respectively. The *entropy* of the source is denoted  $H(P)$ . The *Kullback-Leibler distance*, or *divergence*, between two sources  $P, Q$  over the same alphabet is denoted  $D(P\|Q)$ . We write  $P \ll Q$  if  $Q(x) = 0$  implies  $P(x) = 0$  for all  $x \in \mathcal{X}$ . The set of all p.m.f.’s over  $\mathcal{X}$  is denoted  $\mathcal{P}(\mathcal{X})$ . The *type* of a sequence  $x^n \in \mathcal{X}^n$  is the p.m.f.  $P_{x^n} \in \mathcal{P}(\mathcal{X})$  corresponding to the relative frequency of symbols in  $x^n$ . The set of all possible types of sequences  $x^n$  is denoted  $\mathcal{P}^n(\mathcal{X})$ . The *type class* of any type  $Q \in \mathcal{P}^n(\mathcal{X})$  is the set  $T_Q \stackrel{\text{def}}{=} \{x^n \in \mathcal{X}^n : P_{x^n} = Q\}$ . For  $\varepsilon > 0$ , let  $\mathcal{P}_\varepsilon^n(\mathcal{X}, P) \subseteq \mathcal{P}^n(\mathcal{X})$  be the subset of all types  $Q$  for which  $\|P - Q\|_\infty < \varepsilon$ .

The following facts are well known [22].

**Lemma 2.** *For any type  $Q \in \mathcal{P}^n(\mathcal{X})$  and any  $x^n \in T_Q$ :*

- (i)  $P(x^n) = 2^{-n(D(Q\|P) + H(Q))}$ .
- (ii)  $|\mathcal{P}^n(\mathcal{X})|^{-1} 2^{nH(Q)} \leq |T_Q| \leq 2^{nH(Q)}$ .
- (iii)  $|\mathcal{P}^n(\mathcal{X})| = \binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1} \leq (n+1)^{|\mathcal{X}|}$ .
- (iv) (AEP) For any  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P \left( \bigcup_{Q \in \mathcal{P}_\varepsilon^n(\mathcal{X}, P)} T_Q \right) = 1$$

The Rényi entropy [23] of order  $\alpha$  of a source  $P$  is

$$H_\alpha(P) \stackrel{\text{def}}{=} \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} (P(x))^\alpha$$

**Lemma 3** (From [24]). *The Rényi entropy of order  $\alpha > 1$  admits the following variational characterization:*

$$H_\alpha(P) = \min_{Q \in \mathcal{P}(\mathcal{X})} \left\{ \frac{\alpha}{\alpha-1} D(Q\|P) + H(Q) \right\}$$

For  $0 < \alpha < 1$ , replace the min with a max.

For any two sources  $P, Q$  over the same alphabet  $\mathcal{X}$ , we define

$$\nu(P, Q) \stackrel{\text{def}}{=} \sup_{x \in \mathcal{X}: P(x) > 0} \frac{P(x)}{Q(x)}$$

The following is easy to verify.

**Lemma 4.**  $1 \leq \nu(P, Q) \leq \infty$  with equality in the lower bound if and only if  $P = Q$ , and in the upper bound if and only if  $P \not\ll Q$ .

### C. Encoders and Decoders

An *encoder* is a mapping  $\mathcal{E} : \mathcal{X}^* \mapsto \{0, 1\}^*$  such that for any  $s \in \mathcal{X}^*$ ,  $\mathcal{E}(s)$  is the longest common prefix of the set of bit strings  $\{\mathcal{E}(sx) : x \in \mathcal{X}\}$ . Namely, we are assuming the encoder does not withhold any bits; at any given time, the longest prefix the encoder is certain of is assumed to have already been emitted. This will be referred to as *the integrity property*. Note that the integrity property implies in particular the *consistency property*, namely that  $\mathcal{E}(s) \preceq \mathcal{E}(sx)$ .

An encoder  $\mathcal{E}$  induces a *decoder*, which is a partial mapping  $\mathcal{D}_\mathcal{E} : \{0, 1\}^* \mapsto \mathcal{X}^*$ , defined as follows. For any  $b \in \{0, 1\}^*$ , let

$$\mathcal{E}^{-1}(b) \stackrel{\text{def}}{=} \{s \in \mathcal{X}^* : b \preceq \mathcal{E}(s)\}$$

Then  $\mathcal{D}_\mathcal{E}(b)$  is the longest common prefix of  $\mathcal{E}^{-1}(b)$  if the latter set is not empty, and is otherwise undefined. Note that by definition,  $\mathcal{D}_\mathcal{E}$  does not withhold any symbols, hence satisfies a similar integrity property. Furthermore,  $\mathcal{D}_\mathcal{E}$  is defined not only over the range of  $\mathcal{E}$ , but also on the set of all prefixes thereof; the decoder hence operates without the need to be synced with the source clock. Since a decoder is uniquely defined by an encoder, we shall focus our discussion hereafter on encoders only.

An encoder  $\mathcal{E}$  is associated with a *delay function*, which returns the minimal number of symbols from a given (infinite) suffix that needs to be encoded so that a given prefix can be fully decoded. Formally, the delay function is a mapping  $\delta^\mathcal{E} : \mathcal{X}^* \times \mathcal{X}^\infty \mapsto \mathbb{N} \cup \{0, \infty\}$ , where  $\delta^\mathcal{E}(s, x^\infty)$  is the minimal  $k \in \mathbb{N} \cup \{0\}$  such that  $s \preceq \mathcal{D}_\mathcal{E}(\mathcal{E}(sx^k))$ . If no such  $k$  exists, then  $\delta^\mathcal{E}(s, x^\infty) \stackrel{\text{def}}{=} \infty$ .

The *delay profile* associated with an encoder  $\mathcal{E}$  and a source  $P$  for a given prefix  $s$ , is the following extended-real-valued r.v.:

$$\Delta^\mathcal{E}(s, P) \stackrel{\text{def}}{=} \delta^\mathcal{E}(s, X^\infty)$$

The delay profile associated with an encoder  $\mathcal{E}$  and a source  $P$  is then defined to be

$$\Delta^\mathcal{E}(P) \stackrel{\text{def}}{=} \sup_{s \in \mathcal{X}^*} \Delta^\mathcal{E}(s, P)$$

Next, we define several families of encoders.

1) *Lossless Encoders*: An encoder is said to be *lossless* w.r.t.  $P$  (where  $P$  is omitted when there is no confusion), if

$$\mathbb{P}(\Delta^\mathcal{E}(P) < \infty) = 1,$$

The family of all encoders that are lossless w.r.t.  $P$  is denoted  $\mathfrak{L}(P)$ .

2) *Bounded Expected Delay Encoders*: An encoder is said to admit a *bounded expected delay* w.r.t.  $P$  (where  $P$  is omitted when there is no confusion), if

$$\mathbb{E}(\Delta^\mathcal{E}(P)) < \infty$$

The family of all encoders with bounded expected delay w.r.t.  $P$  is denoted  $\mathfrak{B}(P)$ . Clearly,  $\mathfrak{B}(P) \subset \mathfrak{L}(P)$ .

3) *Delay Constrained Encoders*: An encoder is said to be *delay-constrained*, if

$$\sup_{s \in \mathcal{X}^*, t \in \mathcal{X}^\infty} \delta^\mathcal{E}(s, t) < \infty \quad (1)$$

More specifically, such an encoder is also said to be *d-delay-constrained*, if the supremum above equals  $d$ . The family of *d-constrained* encoders is denoted by  $\mathfrak{C}_d$ .<sup>5</sup> Clearly,  $\mathfrak{C}_d \subset \mathfrak{B}(P)$  for any source  $P$ .

4) *Phrase/Block Constrained Encoders*: An encoder  $\mathcal{E}$  is said to be *phrase-constrained* if  $\mathcal{E} \in \mathfrak{C}_d$  for some  $d$ , and for any  $x^\infty \in \mathcal{X}^\infty$  there exists an index sequence  $\{i_k \in \mathbb{N}\}_{k=1}^\infty$  such that  $0 < i_{k+1} - i_k \leq d + 1$  and

$$\delta^\mathcal{E}(x^{i_k}, x_{i_k+1}^\infty) = 0 \quad (2)$$

In this case we also say the encoder is *d-phrase-constrained*. In the special case where  $i_k = (d+1)k$  for all  $x^\infty \in \mathcal{X}^\infty$ , we say the encoder is *d-block-constrained*. The family of all *d-phrase-constrained* (resp. *d-block-constrained*) encoders is denoted by  $\mathfrak{C}_d^{\text{phrase}}$  (resp.  $\mathfrak{C}_d^{\text{block}}$ ). Clearly,  $\mathfrak{C}_d^{\text{block}} \subset \mathfrak{C}_d^{\text{phrase}} \subset \mathfrak{C}_d$ .

**Remark 1.** Any encoder  $\mathcal{E} \in \mathfrak{C}_d^{\text{block}}$  (resp.  $\mathcal{E} \in \mathfrak{C}_d^{\text{phrase}}$ ) can generally be written as a prefix-dependent concatenation of BV (resp. VB/VV) codes each with block length (resp. maximal phrase length) at most  $d + 1$ . By prefix-dependent here we mean that the code used in the next block (resp. phrase) can generally depend on the source sequence encoded thus far. Note however that for block (resp. phrase) constrained encoders operating over memoryless sources there is no redundancy gain to be reaped by using prefix-dependency, since the entire prefix can already be decoded and hence is irrelevant (in terms of average code-length) to the encoding of the next block (resp. phrase). Hence for memoryless sources, as far as the redundancy-delay tradeoff is concerned, there is no loss of generality in restricting our attention to concatenations of a single fixed BV (resp. VB/VV) code.

Conversely, any BV (resp. VB/VV) code with block length (resp. maximal phrase length)  $k$ , adapted to process infinite source strings via concatenation, is a *d-block-constrained* (resp. *d-phrase-constrained*) code for some  $d \leq k$ . Due to the integrity property requirement, it is generally possible that  $d < k$ , as the base code itself may be a concatenation of

<sup>5</sup>Note that growing dictionary encoders such as the Ziv-Lempel encoder [25] do not belong to this family, as their delay grows unbounded.

shorter codes. This is however clearly redundant, and without loss of generality we can restrict our attention to minimal VB (resp. VB/VV) codes, i.e., codes for which  $k = d$ .

**Remark 2.** Following the previous remark, it is worth mentioning an interesting class of codes known as *plurally parsable (PP) codes* [26], which are a generalization of VB/VV codes. In a nutshell, a PP encoder is defined via a finite phrase dictionary  $\mathfrak{D} \subset \{0, 1\}^*$  and a parsing rule. The dictionary is not a complete code-tree, and hence can induce more than one parsing for some source sequences; in such cases the parsing rule is employed to determine which of the possible parsings will be used. Typically, a greedy parsing rule is employed, looking for the longest match in  $\mathfrak{D}$ . It is interesting to note that while clearly any PP code is delay-constrained, any nontrivial PP code, i.e., one that cannot be essentially translated into a (uniquely parsable) VB/VV code<sup>6</sup>, is not block/phrase constrained, as there are source sequences for which the delay is always positive. For example, using the PP code given by the incomplete code-tree  $\mathfrak{D} = \{0, 000, 1, 111\}$  together with the greedy parsing rule, the delay incurred for the source sequence 001100110011... is always at least 1. Such PP codes hence always look beyond the end of the current phrase.

5) *Interval-Mapping Encoders:* A binary string  $b^k \in \{0, 1\}^k$  is said to represent a binary interval

$$[b^k] \stackrel{\text{def}}{=} [0.b_1b_2, \dots, b_k0, 0.b_1b_2, \dots, b_k1] \subseteq [0, 1]$$

For any set  $A \subset [0, 1]$  we write  $\text{bin}(A)$  to denote the minimal binary interval containing  $A$ , i.e.,

$$\text{bin}(A) \stackrel{\text{def}}{=} \bigcap_{b \in \{0, 1\}^*: A \subseteq [b]} [b]$$

The following lemma is easily observed.

**Lemma 5.** For any  $b, c \in \{0, 1\}^*$ ,

- (i)  $b \leq c \Leftrightarrow [c] \subseteq [b]$ .
- (ii)  $b \not\leq c$  and  $c \not\leq b \Leftrightarrow [b] \cap [c] = \emptyset$ .

Let  $\mathfrak{S} \stackrel{\text{def}}{=} \{[a, b] \mid 0 \leq a < b \leq 1\}$ . An encoder  $\mathcal{E}$  is said to be an *interval-mapping encoder*, if there exists a mapping  $\mathcal{I}^\mathcal{E} : \mathcal{X}^* \mapsto \mathfrak{S}$ , i.e., a mapping of finite source sequences into subintervals of the unit interval, such that the following properties are satisfied:

- (i) *Minimality:*  $[\mathcal{E}(s)] = \text{bin}(\mathcal{I}^\mathcal{E}(s))$  for any  $s \in \mathcal{X}^*$ .
- (ii) *Disjoint nesting:* For all  $s \in \mathcal{X}^*$  and all distinct  $x, y \in \mathcal{X}$ ,

$$\mathcal{I}^\mathcal{E}(sx) \subseteq \mathcal{I}^\mathcal{E}(s), \quad \mathcal{I}^\mathcal{E}(sx) \cap \mathcal{I}^\mathcal{E}(sy) = \emptyset$$

The minimality property means that an interval-mapping encoder emits the bit sequence representing the minimal binary interval containing the interval  $\mathcal{I}^\mathcal{E}(s)$ . It is easily observed that the minimality and disjoint nesting properties together imply

<sup>6</sup>For example, the PP code given by the incomplete code-tree  $\mathfrak{D} = \{0, 00, 1\}$  together with the greedy parsing rule, can essentially be thought of as a uniquely parsable code given by the complete code-tree  $\mathfrak{D} = \{00, 01, 1\}$ , in the sense that the parsing induced by the former is a refinement of the parsing induced by the latter.

the integrity property. The family of interval mapping encoders is denoted by  $\mathfrak{I}$ .

Let  $<$  be any order of  $\mathcal{X}$ . A special case of an interval-mapping encoder is an *arithmetic encoder w.r.t. the order  $<$  matched to a source  $P$* , which is defined as follows:

$$\begin{aligned} f_1(x) &\stackrel{\text{def}}{=} \sum_{y < x} P(y) \\ f_n(x^n) &\stackrel{\text{def}}{=} f_{n-1}(x^{n-1}) + f_1(x_n)P(x^{n-1}) \\ \mathcal{I}^\mathcal{E}(x^n) &\stackrel{\text{def}}{=} [f_n(x^n), f_n(x^n) + P(x^n)] \end{aligned}$$

We omit the reference to a specific order  $<$  when there is no confusion, or when the statement holds for any order.

6) *Generalized Interval-Mapping Encoders:* Let  $\mathfrak{S}^*$  be the set of all finite disjoint unions of subintervals from  $\mathfrak{S}$ . An encoder  $\mathcal{E}$  is said to be a *generalized interval-mapping encoder* if there exists a mapping  $\mathcal{I}^\mathcal{E} : \mathcal{X}^* \mapsto \mathfrak{S}^*$  satisfying the minimality and disjoint nesting properties above. The family of generalized interval-mapping encoders is denoted by  $\mathfrak{I}^*$ . Clearly,  $\mathfrak{I} \subset \mathfrak{I}^*$ .

The following lemma shows that any  $d$ -delay-constrained encoder admits a generalized interval-mapping representation.

**Lemma 6.** Let  $\mathcal{E} \in \mathfrak{C}_d$ . Then  $\mathcal{E}$  can be represented as a generalized interval-mapping encoder with

$$\mathcal{I}^\mathcal{E}(s) = \bigcup_{x^d \in \mathcal{X}^d} [\mathcal{E}(sx^d)] \quad (3)$$

Hence,  $\mathfrak{C}_d \subset \mathfrak{I}^*$ .

*Proof:* See the Appendix. ■

**Remark 3.** The representation in (3) is a finite union of (possibly overlapping) binary intervals. It is worth noting that an arithmetic encoder matched to a source cannot generally be written that way, as some of its intervals may only be written as an infinite union of binary intervals. This sits well with the fact that generally, an (idealized) arithmetic encoder has an unbounded delay.

#### D. Redundancy

The (per symbol) expected codelength at time  $n$  associated with an encoder  $\mathcal{E}$  and a memoryless source  $P$  is

$$\bar{L}_n^\mathcal{E}(P) \stackrel{\text{def}}{=} n^{-1} \mathbb{E}[\mathcal{E}(X^n)] \quad (4)$$

where  $X^n \sim P^n$ . The (per symbol) expected redundancy at time  $n$  associated with an encoder  $\mathcal{E}$  and a memoryless source  $P$  is the gap between the expected codelength and the entropy after  $n$  symbols have been encoded, i.e.,

$$\mathfrak{R}_n^\mathcal{E}(P) \stackrel{\text{def}}{=} \bar{L}_n^\mathcal{E} - H(P)$$

The corresponding *sup-redundancy* and *inf-redundancy* are defined as

$$\overline{\mathfrak{R}}^\mathcal{E}(P) \stackrel{\text{def}}{=} \limsup_{n \rightarrow \infty} \mathfrak{R}_n^\mathcal{E}(P), \quad \underline{\mathfrak{R}}^\mathcal{E}(P) \stackrel{\text{def}}{=} \liminf_{n \rightarrow \infty} \mathfrak{R}_n^\mathcal{E}(P)$$

Let us define some useful quantities pertaining to generalized interval-mapping encoders, that will enable us to bound their redundancy in relatively simple terms. A generalized

interval-mapping encoder  $\mathcal{E}$  induces a measure over  $\mathcal{X}^n$ , defined by

$$\mu_n^\mathcal{E}(x^n) \stackrel{\text{def}}{=} |\mathcal{I}^\mathcal{E}(x^n)|$$

and a conditional induced measure, defined as

$$\mu_k^\mathcal{E}(x^k|x^n) \stackrel{\text{def}}{=} \frac{\mu_{n+k}^\mathcal{E}(x^n x^k)}{\mu_n^\mathcal{E}(x^n)}$$

Define:

$$R_n^\mathcal{E}(P) \stackrel{\text{def}}{=} \frac{1}{n} D(P^n \| \mu_n^\mathcal{E})$$

and let

$$r_d(x^n) = D(P^d \| \mu_d^\mathcal{E}(\cdot|x^n))$$

be the  $d$ -instantaneous redundancy.

**Remark 4.** Note that  $\mu_n^\mathcal{E}$  and  $\mu_k^\mathcal{E}(\cdot|x^n)$  are not necessarily probability distributions, as they may sum to less than unity. However, clearly it still holds that  $R_n^\mathcal{E}(P) \geq 0, r_d(x^n) \geq 0$ .

The next lemma relates the interval-based notions of redundancy defined above, to the actual operational definition of redundancy of the associated generalized interval-mapping encoders. This correspondence will allow us to think of intervals instead of bits, and will play a central role in the sequel.

**Lemma 7.** The following relations hold:

(i) For any  $\mathcal{E} \in \mathfrak{J}^*$ ,

$$\mathfrak{R}_n^\mathcal{E}(P) \leq R_n^\mathcal{E}(P)$$

(ii) For any  $\mathcal{E} \in \mathfrak{C}_d$ , there exists a generalized interval-mapping representation  $\mathcal{I}^\mathcal{E}$  (e.g., the one in Lemma 6) such that

$$\begin{aligned} \mathfrak{R}_n^\mathcal{E}(P) &\geq \left(\frac{n+d}{n}\right) R_{n+d}^\mathcal{E}(P) + \frac{d}{n} H(P) \\ \underline{\mathfrak{R}}^\mathcal{E}(P) &= \liminf_{n \rightarrow \infty} \frac{1}{nd} \sum_{k=1}^n \mathbb{E}(r_d(X^k)) \end{aligned}$$

*Proof:* See the Appendix. ■

One would naturally be interested in the redundancy performance that can be guaranteed by employing encoders of different classes. In general, the expected redundancy  $\mathfrak{R}_n^\mathcal{E}$  of an encoder  $\mathcal{E}$  can be negative for some, or even all  $n$ . However, the sup and inf-redundancy are nonnegative for all lossless encoders, and bounds in the  $d$ -block/phrase constrained cases are known.

**Lemma 8.** The following statements hold<sup>7</sup>:

(i) For any source  $P$

$$\begin{aligned} \inf_{\mathcal{E} \in \mathfrak{L}(P)} \overline{\mathfrak{R}}^\mathcal{E}(P) &= \inf_{\mathcal{E} \in \mathfrak{B}(P)} \overline{\mathfrak{R}}^\mathcal{E}(P) = \inf_{\mathcal{E} \in \mathfrak{L}(P)} \underline{\mathfrak{R}}^\mathcal{E}(P) \\ &= \inf_{\mathcal{E} \in \mathfrak{B}(P)} \underline{\mathfrak{R}}^\mathcal{E}(P) = 0 \end{aligned}$$

<sup>7</sup>Recall that  $f(d) = O(g(d)) \Rightarrow \limsup_{d \rightarrow \infty} \left| \frac{f(d)}{g(d)} \right| < \infty$ , and  $f(d) = \Omega(g(d)) \Rightarrow \liminf_{d \rightarrow \infty} \left| \frac{f(d)}{g(d)} \right| > 0$

(ii) (From [1], [7], [6]) For any source

$$\inf_{\mathcal{E} \in \mathfrak{C}_d^{\text{block}}} \overline{\mathfrak{R}}^\mathcal{E}(P) = O(d^{-1}), \quad \inf_{\mathcal{E} \in \mathfrak{C}_d^{\text{phrase}}} \overline{\mathfrak{R}}^\mathcal{E}(P) = O(d^{-\frac{5}{3}})$$

(iii) (From [5], [6]) For almost all sources,

$$\begin{aligned} \inf_{\mathcal{E} \in \mathfrak{C}_d^{\text{block}}} \underline{\mathfrak{R}}^\mathcal{E}(P) &= \Omega(d^{-1}) \\ \inf_{\mathcal{E} \in \mathfrak{C}_d^{\text{phrase}}} \underline{\mathfrak{R}}^\mathcal{E}(P) &= \Omega(d^{-2|\mathcal{X}|-1-\varepsilon}) \end{aligned}$$

where  $\varepsilon > 0$ .

We see that employing block/phrase-constrained codes for compression under a strict delay constraint, the redundancy decays at best polynomially with the delay constraint<sup>8</sup>. As we shall see, the redundancy can be made to decay exponentially with the delay, if the more general family of delay-constrained encoders is used. This reveals a fundamental difference between block/phrase length and delay in lossless source coding.

The following lemma shows that for an optimal  $d$ -delay-constrained encoder, the inf-redundancy and sup-redundancy coincide.

**Lemma 9.** For any source  $P$ ,

$$\inf_{\mathcal{E} \in \mathfrak{C}_d} \overline{\mathfrak{R}}^\mathcal{E}(P) = \inf_{\mathcal{E} \in \mathfrak{C}_d} \underline{\mathfrak{R}}^\mathcal{E}(P) \stackrel{\text{def}}{=} \mathfrak{R}(P, d)$$

*Proof:* See the Appendix. ■

Accordingly,  $\mathfrak{R}(P, d)$  defined above is called the *redundancy-delay function* associated with the source  $P$ . The corresponding inf-redundancy-delay and sup-redundancy-delay exponents associated with  $P$  can now be defined:

$$\begin{aligned} \overline{E}(P) &= \limsup_{d \rightarrow \infty} -\frac{1}{d} \log \mathfrak{R}(P, d) \\ \underline{E}(P) &= \liminf_{d \rightarrow \infty} -\frac{1}{d} \log \mathfrak{R}(P, d) \end{aligned}$$

Our main goal in this paper is to characterize  $\mathfrak{R}(P, d)$ ,  $\overline{E}(P)$  and  $\underline{E}(P)$ .

### III. THE DELAY PROFILE OF ARITHMETIC CODING

Consider a case where a source  $P$  is encoded by a mismatched arithmetic encoder, namely where the encoder's interval lengths match a different source  $Q$  (see also Subsection II-C). Note that we can always assume that  $P \ll Q$ , as otherwise the mismatched encoder is not well defined for all input symbols. In the next theorem we upper bound the probability that the corresponding delay profile exceeds a given threshold. This result will serve as a tool in the next section, where we lower bound the redundancy-delay exponent.

**Theorem 1.** Suppose a source  $P \in \mathcal{P}(\mathcal{X})$  is encoded using an arithmetic encoder  $\mathcal{E}$  matched to a source  $Q \in \mathcal{P}(\mathcal{X})$ , where  $P \ll Q$ . Then

$$\begin{aligned} \mathbb{P}(\Delta^\mathcal{E}(P) > d) &\leq 2p_{\max}^d \left( d \log \left( \frac{\nu(P, Q)}{p_{\max}} \right) + \kappa \right) \\ &\quad + 2q_{\max}^d (\nu(P, Q))^d \end{aligned} \quad (5)$$

<sup>8</sup>This is in fact true even under the weaker expected delay constraint.

where  $\kappa = \log\left(\frac{\sqrt{2e}}{\log e}\right) \approx 1.4139 \dots$

An outline of the proof is given in Section III-A. The full proof is given in Section III-C.

**Corollary 1.** *Let  $\mathcal{E}$  be an arithmetic encoder matched to a source  $Q \in \mathcal{P}(\mathcal{X})$ , where  $P \ll Q$ . For any source  $P \in \mathcal{P}(\mathcal{X})$ , if*

$$q_{\max} \cdot \nu(P, Q) < 1$$

*then the delay profile bound (5) is exponentially decaying with  $d$ , hence the expected delay is finite, i.e.,  $\mathcal{E} \in \mathfrak{B}(P)$ . This specifically holds for all non-deterministic  $P = Q$ .*

**Corollary 2.** *Suppose the source  $P$  is encoded using the arithmetic encoder matched to the source. Then*

$$\mathbb{P}(\Delta^{\mathcal{E}}(P) > d) \leq 2p_{\max}^d (d \log(1/p_{\max}) + \kappa + 1)$$

**Remark 5.** *A bound on the moment-generating function for matched arithmetic coding, and a corresponding exponential bound on the delay's tail distribution, were originally observed in [19], [15]. However, these bounds depend on both  $p_{\min}$  and  $p_{\max}$ , and can therefore be arbitrarily loose. For the tail distribution, a bound depending only on  $p_{\max}$  was originally obtained by the authors in [16], where it was also shown how the proof of [19], [15] can be tweaked to remove the dependency on  $p_{\min}$ . The bound obtained here is tighter than both.*

**Remark 6.** *The bound in Theorem 1 can be further tightened by observing that specific orders of the alphabet  $\mathcal{X}$  are better than others in terms of the bounding technique used here. We do not pursue this direction, since we need an order-independent bound in the sequel.*

#### A. Proof Outline

Recall the definitions of an interval-mapping encoder and of an arithmetic encoder in particular, given in Subsection II-C. At time  $n$ , the sequence  $x^n$  has been encoded into  $\mathcal{I}^{\mathcal{E}}(x^n)$ , and the decoder is so far aware only of the interval  $\text{bin}(\mathcal{I}^{\mathcal{E}}(x^n))$ , namely the minimal binary interval containing  $\mathcal{I}^{\mathcal{E}}(x^n)$ . Thus the decoder is able to decode  $x^m$ , where  $m$  is maximal such that  $\text{bin}(\mathcal{I}^{\mathcal{E}}(x^n)) \subseteq \mathcal{I}^{\mathcal{E}}(x^m)$ . Of course,  $m \leq n$  where the inequality is generally strict. After  $d$  more source letters are fed to the encoder,  $x^{n+d}$  is encoded into  $\mathcal{I}^{\mathcal{E}}(x^{n+d})$ , and the entire sequence  $x^n$  can be decoded at time  $n + d$  if and only if<sup>9</sup>

$$\text{bin}(\mathcal{I}^{\mathcal{E}}(x^{n+d})) \subseteq \mathcal{I}^{\mathcal{E}}(x^n). \quad (6)$$

Now, consider the midpoint of  $\text{bin}(\mathcal{I}^{\mathcal{E}}(x^n))$  which by the minimality property (see Subsection II-C) is always contained in  $\mathcal{I}^{\mathcal{E}}(x^n)$ . If that midpoint is contained in  $\mathcal{I}^{\mathcal{E}}(x^{n+d})$  (but not as a left edge), then condition (6) cannot be satisfied; In fact, in this case the encoder cannot yield even one further bit. This observation can be generalized to a set of points which, if contained in  $\mathcal{I}^{\mathcal{E}}(x^{n+d})$ ,  $x^n$  cannot be completely decoded. For each of these points the encoder outputs a number of bits which may enable the decoder to produce source symbols, but

not enough to fully decode  $x^n$ . The encoding and decoding delays are therefore treated here simultaneously, rather than separately as in [15].

**Remark 7.** *When  $Q \not\ll P$  there are “holes” in the interval-mapping, namely intervals corresponding to symbols where  $Q(x) > 0$  but  $P(x) = 0$ . In this case,  $x^n$  can be decoded at time  $n + d$  if and only if  $\text{bin}(\mathcal{I}^{\mathcal{E}}(x^{n+d})) \cap \mathcal{I}^{\mathcal{E}}(y^n) = \emptyset$  for any  $y^n \neq x^n$ . Hence condition (6) is necessary and sufficient if  $Q \ll P$ , and only sufficient otherwise. This point is important to note since the case where  $Q \not\ll P$  appears in the sequel.*

After having identified the above set of *forbidden points*, we clearly need to analyze the probability of avoiding them within the next  $d$  instances. Loosely speaking, for an arithmetic encoder matched to the source  $P$ , the maximal symbol probability  $p_{\max}$  represents the “crudest resolution”, or the “lowest rate” by which we shrink our intervals, hence intuitively dictates our ability to avoid hitting forbidden points. Indeed, the probability that the encoder avoids these points is roughly  $p_{\max}^d$ . For a mismatched encoder, we get a similar expression involving  $p_{\max}^d, q_{\max}^d$  and  $\nu(P, Q)$  as a measure of the mismatch between the encoder and the source.

#### B. The Forbidden Points Notion

We now introduce some notations and prove three lemmas, required for the proof of Theorem 1. Let  $I = [a, b) \subseteq [0, 1)$  be some interval, and  $p$  some point in that interval. We say that  $p$  is *strictly contained* in  $I$  if  $p \in (a, b)$ . We define the *left-adjacent* of  $p$  w.r.t.  $I$  to be

$$\ell_I(p) \stackrel{\text{def}}{=} \min \{x \in [a, p) : \exists k \in \mathbb{Z}^+, x = p - 2^{-k}\}$$

and the *t-left-adjacent* of  $p$  w.r.t.  $I$  as

$$\ell_I^{(t)}(p) \stackrel{\text{def}}{=} \overbrace{(\ell_I \circ \ell_I \circ \dots \circ \ell_I)}^t(p), \quad \ell_I^{(0)}(p) \stackrel{\text{def}}{=} p$$

Notice that  $\ell_I^{(t)}(p) \rightarrow a$  monotonically with  $t$ . We also define the *right-adjacent* of  $p$  w.r.t.  $I$  to be

$$r_I(p) \stackrel{\text{def}}{=} \max \{x \in (p, b) : \exists k \in \mathbb{Z}^+, x = p + 2^{-k}\}$$

and  $r_I^{(t)}(p)$  as the *t-right-adjacent* of  $p$  w.r.t.  $[a, b)$  similarly, where now  $r_I^{(t)}(p) \rightarrow b$  monotonically. For any  $\delta < b - a$ , the *adjacent  $\delta$ -set* of  $p$  w.r.t.  $I$  is defined as the set of all adjacents that are not “too close” to the edges of  $I$ :

$$S_{\delta}(I, p) \stackrel{\text{def}}{=} \{x \in [a + \delta, b - \delta) : \exists t \in \mathbb{Z}^+ \cup \{0\}, \\ x = \ell_I^{(t)}(p) \vee x = r_I^{(t)}(p)\}$$

Notice that for  $\delta > p - a$  this set may contain only right-adjacents, for  $\delta > b - p$  only left-adjacents, for  $\delta > \frac{b-a}{2}$  it is empty, and for  $\delta = 0$  it may be infinite.

**Lemma 10.** *The size of  $S_{\delta}(I, p)$  is upper bounded by*

$$|S_{\delta}(I, p)| \leq 1 + 2 \log \frac{|I|}{\delta} \quad (7)$$

*Proof:* See the Appendix. ■

For an interval  $I$ , let  $m(I)$  denote the midpoint of  $\text{bin}(I)$ . Note that  $m(I) \in I$ , by definition of  $\text{bin}(I)$  as the minimal

<sup>9</sup>Here we are further assuming that  $Q \ll P$ , see Remark 7.

binary interval containing  $I$ . In what follows, we will be specifically interested in the adjacent  $\delta$ -set of  $m(I)$  w.r.t.  $I$ . We therefore suppress the dependence on  $m(I)$  and write

$$S_\delta(I) \stackrel{\text{def}}{=} S_\delta(I, m(I))$$

In particular, the set  $S_0(I)$  will be referred to as the *forbidden points* of  $I$ . The forbidden points play a central role in the sequel, for the following reason:

**Lemma 11.** *Condition (6) is satisfied if and only if  $\mathcal{I}^\mathcal{E}(x^{n+d})$  does not contain forbidden points of  $\mathcal{I}^\mathcal{E}(x^n)$ , i.e.,*

$$\mathcal{I}^\mathcal{E}(x^{n+d}) \cap S_0(\mathcal{I}^\mathcal{E}(x^n)) = \emptyset$$

*Proof:* Write  $m = m(\mathcal{I}^\mathcal{E}(x^n))$  for short. As already discussed, if  $m$  is strictly contained in  $\mathcal{I}^\mathcal{E}(x^{n+d})$  then (6) is not satisfied. Otherwise, assume  $\mathcal{I}^\mathcal{E}(x^{n+d})$  lies to the left of  $m$ . Clearly, if  $\mathcal{I}^\mathcal{E}(x^{n+d}) \subseteq [\ell(m), m)$ , then  $\text{bin}(\mathcal{I}^\mathcal{E}(x^{n+d})) \subseteq [\ell(m), m)$  as well, hence (6) is satisfied. However, if  $\ell(m)$  is strictly contained in  $\mathcal{I}^\mathcal{E}(x^{n+d})$  then  $\text{bin}(\mathcal{I}^\mathcal{E}(x^{n+d}))$  must be the left half of  $\text{bin}(\mathcal{I}^\mathcal{E}(x^n))$ , which by minimality cannot be a subinterval of  $\mathcal{I}^\mathcal{E}(x^n)$ , hence (6) is not satisfied. The same rationale also applies to  $r(m)$ . The lemma follows by iterating the argument. ■

### C. Proof of Theorem 1

The probability that the delay  $\Delta^\mathcal{E}(x^n, P)$  is larger than  $d$  is equal to (or upper bounded by, when  $Q \not\ll P$ , see Remark 7) the probability that (6) is not satisfied. By Lemma 11, this in turn equals the probability that  $\mathcal{I}^\mathcal{E}(X^{n+d})$  contains none of the forbidden points of  $\mathcal{I}^\mathcal{E}(x^n)$ . To get a handle on this latter probability, the following lemma is found useful.

**Lemma 12.** *Suppose a source  $P$  is encoded using an arithmetic encoder  $\mathcal{E}$  matched to a source  $Q$ , where  $P \ll Q$ , and let  $p_{\max}, q_{\max}$  be the corresponding maximal symbol probabilities. Then for any  $a \in \mathcal{I}^\mathcal{E}(x^n)$ ,*

$$\mathbb{P}(a \in \mathcal{I}^\mathcal{E}(X^{n+d}) | X^n = x^n) \leq p_{\max}^d$$

and for any interval  $J \subseteq \mathcal{I}^\mathcal{E}(x^n)$  sharing an endpoint with  $\mathcal{I}^\mathcal{E}(x^n)$ ,

$$\begin{aligned} \mathbb{P}(J \cap \mathcal{I}^\mathcal{E}(X^{n+d}) \neq \emptyset | X^n = x^n) \\ \leq \left( \frac{|J|}{|\mathcal{I}^\mathcal{E}(x^n)|} + q_{\max}^d \right) (\nu(P, Q))^d \end{aligned}$$

*Proof:* The set  $\{\mathcal{I}^\mathcal{E}(x^n y^d) : y^d \in \mathcal{X}^d\}$  is a partition of  $\mathcal{I}^\mathcal{E}(x^n)$  into intervals, and  $a$  belongs to a single interval in the partition. Therefore,

$$\begin{aligned} \mathbb{P}(a \in \mathcal{I}^\mathcal{E}(X^{n+d}) | X^n = x^n) \\ \leq \max_{y^d \in \mathcal{X}^d} \mathbb{P}(X_{n+1}^{n+d} = y^d | X^n = x^n) = p_{\max}^d \end{aligned} \quad (8)$$

establishing the first assertion. For the second assertion, write:

$$\begin{aligned} \mathbb{P}(J \cap \mathcal{I}^\mathcal{E}(X^{n+d}) \neq \emptyset | X^n = x^n) &\leq \sum_{y^d: J \cap \mathcal{I}^\mathcal{E}(x^n y^d) \neq \emptyset} P(y^d) \\ &\leq \sum_{y^d: J \cap \mathcal{I}^\mathcal{E}(x^n y^d) \neq \emptyset} Q(y^d) \cdot (\nu(P, Q))^d \\ &= (\nu(P, Q))^d \sum_{y^d: J \cap \mathcal{I}^\mathcal{E}(x^n y^d) \neq \emptyset} \mu_d^\mathcal{E}(y^d | x^n) \\ &\leq \left( \frac{|J|}{|\mathcal{I}^\mathcal{E}(x^n)|} + q_{\max}^d \right) (\nu(P, Q))^d \end{aligned} \quad (9)$$

where we have used the fact that  $\max_{y^d} \mu_d^\mathcal{E}(y^d | x^n) = q_{\max}^d$ . ■

Write  $S_\delta = S_\delta(\mathcal{I}^\mathcal{E}(x^n))$  for short. Note that  $S_\delta \subseteq S_0$ , and that  $S_0 \setminus S_\delta$  is contained in two intervals of length  $\delta$  both sharing an edge with  $\mathcal{I}^\mathcal{E}(x^n)$ . For any  $\delta > 0$ , the delay's tail probability is bounded as follows:

$$\begin{aligned} \mathbb{P}(\Delta^\mathcal{E}(x^n, P) > d) &\stackrel{(a)}{\leq} \mathbb{P}(\text{bin}(\mathcal{I}^\mathcal{E}(X^{n+d})) \not\subseteq \mathcal{I}^\mathcal{E}(x^n) | X^n = x^n) \\ &\stackrel{(b)}{=} \mathbb{P}(S_0 \cap \mathcal{I}^\mathcal{E}(X^{n+d}) \neq \emptyset | X^n = x^n) \\ &\stackrel{(c)}{\leq} \mathbb{P}((S_0 \setminus S_\delta) \cap \mathcal{I}^\mathcal{E}(X^{n+d}) \neq \emptyset | X^n = x^n) \\ &\quad + \mathbb{P}(S_\delta \cap \mathcal{I}^\mathcal{E}(X^{n+d}) \neq \emptyset | X^n = x^n) \\ &\stackrel{(d)}{\leq} 2 \left( \frac{\delta}{|\mathcal{I}^\mathcal{E}(x^n)|} + q_{\max}^d \right) (\nu(P, Q))^d \\ &\quad + p_{\max}^d |S_\delta| \\ &\stackrel{(e)}{\leq} 2 \left( \frac{\delta}{|\mathcal{I}^\mathcal{E}(x^n)|} + q_{\max}^d \right) (\nu(P, Q))^d \\ &\quad + p_{\max}^d \left( 1 + 2 \log \frac{|\mathcal{I}^\mathcal{E}(x^n)|}{\delta} \right) \end{aligned} \quad (10)$$

The transitions are justified as follows:

- (a) Condition (6) is sufficient, see discussion in Subsection III-A. In most cases this would be an equality, as condition (6) would be also necessary, see Remark 7.
- (b) Lemma 11.
- (c) Union bound over  $S_0 = S_\delta \cup (S_0 \setminus S_\delta)$ .
- (d) Lemma 12, together with a union bound over the finite number of elements in  $S_0 \setminus S_\delta$ .
- (e) Lemma 10.

Taking the derivative of the right-hand-side of (10) w.r.t.  $\delta$  we find that  $\delta = \log e \left( \frac{p_{\max}}{\nu(P, Q)} \right)^d |\mathcal{I}^\mathcal{E}(x^n)|$  minimizes the bound. Substituting into (10) and noting that the bound is independent of  $x^n$ , (5) is proved<sup>10</sup>.

## IV. A LOWER BOUND FOR $\underline{E}(P)$

In this section we use the delay's tail distribution mentioned in the previous section, to derive an upper bound for the redundancy–delay function, and hence a lower bound on the inf–redundancy–delay exponent, via a specific arithmetic coding scheme. We emphasize that unlike [21], the presented scheme is error free, hence there is zero probability of buffer

<sup>10</sup>Observe that (10) holds even if  $\delta > |\mathcal{I}^\mathcal{E}(x^n)|$ , in which case our bound becomes trivial.

overflow. Moreover, our figure of merit is the delay in source symbols vs. the redundancy in encoded bits per symbol. ■

#### A. A Finite Delay Result

**Theorem 2.** *The redundancy–delay function for a source  $P$  is upper bounded by*

$$\mathfrak{R}(P, d) \leq 2p_{\max}^{d-c(p_{\max})} \left( (d - c(p_{\max})) \log(2/p_{\max}) + 1 + \kappa \right)^2 \quad (11)$$

where

$$c(x) = \begin{cases} 0 & x < \frac{1}{16} \\ 2 \left\lfloor \frac{1}{\log(2/x)} \right\rfloor - 1 & o.w. \end{cases}$$

**Corollary 3.** *The inf–redundancy–delay exponent for a source  $P$  is lower bounded by*

$$\underline{E}(P) \geq \log(1/p_{\max})$$

*Proof:* Let us first describe the high-level idea behind the proof. We extend the source's alphabet by adding two *fictitious symbols*, and then encode the source using a slightly mismatched arithmetic encoder. The encoder keeps track of the decoding delay, and whenever the delay reaches  $d + 1$ , it inserts a fictitious symbol that nullifies the delay. There are three key points: 1) There exists a mapping such that there is always at least one fictitious symbol whose interval contains no forbidden points, 2) The length assigned to the fictitious symbols can be made very small, and 3) The probability of insertion, bounded via Theorem 1, is also very small.

For any interval  $I = [a, b)$ , let

$$\varphi_I(\lambda) \stackrel{\text{def}}{=} (1 - \lambda)a + \lambda b$$

and define the two disjoint subintervals

$$I_L \stackrel{\text{def}}{=} (\varphi_I(3/8), \varphi_I(1/2)), \quad I_R \stackrel{\text{def}}{=} (\varphi_I(1/2), \varphi_I(5/8))$$

The first key point is established in the following Lemma.

**Lemma 13.** *For any interval  $I \subseteq [0, 1)$ , either  $I_L \cap S_0(I) = \emptyset$  or  $I_R \cap S_0(I) = \emptyset$ .*

*Proof of Lemma 13:* Write  $m = m(\mathcal{I}^{\mathcal{E}}(x^n))$  for short. Without loss of generality, assume that  $m \leq \varphi_I(1/2)$ . There are two cases:

- (1)  $m \leq \varphi_I(3/8)$ : It is easily verified that the right adjacent of  $m$  satisfies  $r(m) > \varphi_I(1/2)$ , as otherwise

$$m + 2(r(m) - m) \in I$$

contradicting the maximality in the definition of the right adjacent. Therefore in this case  $I_L$  contains no forbidden points of  $I$ .

- (2)  $m > \varphi_I(3/8)$ : By our assumption  $m < \varphi_n(1/2)$ , hence

$$r(m) - m \geq \frac{\varphi_I(1) - \varphi_I(1/2)}{2}$$

Rewriting, we have

$$r(m) \geq m + \frac{\varphi_I(1) - \varphi_I(1/2)}{2} \geq \varphi_I(5/8)$$

and therefore  $I_R$  contains no forbidden points.

Returning to the proof of Theorem 2, define an extended alphabet  $\mathcal{X}^+ = \mathcal{X} \cup \{x_L, x_R\}$  where  $x_L, x_R$  are two *fictitious symbols*. Let  $P^+ \in \mathcal{P}(\mathcal{X}^+)$  be the corresponding extension of the source  $P$  to  $\mathcal{X}^+$ , assigning zero probability to the fictitious symbols. For  $0 < \varepsilon < p_{\max}$ , let  $P_{\varepsilon}^+ \in \mathcal{P}(\mathcal{X}^+)$  be a source with the following symbol probabilities:

$$P_{\varepsilon}^+(x) = \begin{cases} (1 - 2\varepsilon)P(x) & x \in \mathcal{X} \\ \varepsilon & x \in \{x_L, x_R\} \end{cases}$$

Clearly,  $\max P_{\varepsilon}^+(x) = (1 - 2\varepsilon)p_{\max} < \frac{1}{16}$  and  $\nu(P^+, P_{\varepsilon}^+) = \frac{1}{1-2\varepsilon}$ . Let  $<$  be any order of  $\mathcal{X}$ . Assuming  $P_{\varepsilon}^+(x) < \frac{1}{16}$  for all  $x \in \mathcal{X}^+$ , and since  $|I_L| = |I_R| = |I|/8$ , then it is easy to see there exists a order  $<^+$  of  $\mathcal{X}^+$  that preserves  $<$  over  $\mathcal{X}$ , such that the arithmetic encoder  $\mathcal{E}$  w.r.t.  $<^+$  matched to  $P_{\varepsilon}^+$  has the fictitious symbols  $x_L, x_R$  mapped into intervals contained in  $\mathcal{I}^{\mathcal{E}}(x^n)_L$  and  $\mathcal{I}^{\mathcal{E}}(x^n)_R$ , respectively. If the condition on  $p_{\max}$  is not satisfied, then we can always aggregate a few symbols into a super-symbol, so that the maximal product probability satisfies the required condition (the effect of this aggregation on the delay is treated later on). To encode the source  $P^+$ , let us now use the arithmetic encoder for  $P_{\varepsilon}^+$  above together with the following fictitious symbol insertion algorithm: The encoder keeps track of the decoding delay by emulating the decoder. Whenever this delay reaches  $d + 1$ , the encoder finds which one of  $\mathcal{I}^{\mathcal{E}}(x^n)_L$  or  $\mathcal{I}^{\mathcal{E}}(x^n)_R$  contains no forbidden point as guaranteed by Lemma 13, and inserts the corresponding fictitious symbol  $x_L$  or  $x_R$  respectively, hence nullifying the decoding delay. This way, the decoding delay never exceeds  $d$  and no errors are incurred.

We now bound the redundancy incurred by the encoder  $\mathcal{E}' \in \mathfrak{C}_d$  described above. There are two different sources of redundancy. The first is due to the mismatch between  $P^+$  and  $P_{\varepsilon}^+$ , and the second is due to the coding of the inserted fictitious symbol. At each time  $k > d$ , the probability  $w_k$  for an insertion can be bounded via Theorem 1:

$$\begin{aligned} w_k &= \mathbb{P}(\Delta^{\mathcal{E}'}(X^{k-d}, P) > d) \leq \mathbb{P}(\Delta^{\mathcal{E}'}(P) > d) \\ &\leq 2p_{\max}^d \left( d \log \left( \frac{1}{(1 - 2\varepsilon)p_{\max}} \right) + \kappa \right) \\ &\quad + 2(1 - 2\varepsilon)^d p_{\max}^d (1 - 2\varepsilon)^{-d} \\ &= 2p_{\max}^d \left( d \log \left( \frac{1}{(1 - 2\varepsilon)p_{\max}} \right) + \kappa + 1 \right) \end{aligned} \quad (12)$$

Now, let  $P^{+n}$  be the  $n$ -product of  $P^+$ , and write

$$\begin{aligned} \mathfrak{R}_n^{\mathcal{E}'}(P) &= \mathfrak{R}_n^{\mathcal{E}'}(P^+) \stackrel{(a)}{\leq} R_n^{\mathcal{E}'}(P^+) = \frac{1}{n} D(P^{+n} \| \mu_n^{\mathcal{E}'}) \\ &\stackrel{(b)}{=} \frac{1}{n} \sum_{k=1}^n \mathbb{E} \left( D(P^+ \| \mu_1^{\mathcal{E}'}(\cdot | X^{k-1})) \right) \\ &\stackrel{(c)}{=} D(P^+ \| P_{\varepsilon}^+) + \frac{1}{n} \log \frac{1}{\varepsilon} \sum_{k=1}^n w_k \\ &\stackrel{(d)}{\leq} 2 \log \left( \frac{1}{\varepsilon} \right) p_{\max}^d \left( d \log \left( \frac{1}{(1 - 2\varepsilon)p_{\max}} \right) + \kappa + 1 \right) \\ &\quad + \log \frac{1}{1 - 2\varepsilon} \end{aligned}$$



$$\stackrel{(e)}{\leq} 2 \log\left(\frac{1}{\varepsilon}\right) p_{\max}^d \left(2d \log\left(\frac{2}{p_{\max}}\right) + \kappa + 1\right) + 4\varepsilon$$

The transitions are justified as follows:

- (a) Lemma 7.
- (b) The chain rule for the divergence, and the fact that  $P^{n+}$  is a product (memoryless) distribution.
- (c) Given  $X^{k-1}$ ,  $\mu_1^\varepsilon$  follows  $P_\varepsilon^+$  with an extra multiplication by  $\varepsilon$  if and only if  $X^{k-1}$  is such that there is an insertion. Hence the expected divergence given  $X^{k-1}$  always yields the term  $D(P^+ \| P_\varepsilon^+)$ , and an extra  $\log 1/\varepsilon$  multiplied by the probability of an insertion  $w_k$ .
- (d) The bound for  $w_k$  given in (12), and  $D(P^+ \| P_\varepsilon^+) = \log \frac{1}{1-2\varepsilon}$ .
- (e)  $\log \frac{1}{1-2\varepsilon} \leq 4\varepsilon$  for  $0 < \varepsilon < \frac{1}{16}$ .

Setting  $\varepsilon = p_{\max}^d$ , we get:

$$\begin{aligned} \mathfrak{R}_n^{\varepsilon'}(P) &\leq 2p_{\max}^d \left(d \log\left(\frac{2}{p_{\max}}\right) + \kappa + 1\right) d \log \frac{1}{p_{\max}} + 4p_{\max}^d \\ &\leq 2p_{\max}^d \left(d \log\left(\frac{2}{p_{\max}}\right) + \kappa + 1\right)^2 \end{aligned} \quad (13)$$

Finally, we address the case where  $p_{\max} > \frac{1}{16}$ . As mentioned before, we aggregate a minimal number of source symbols  $k$  into a super-symbol, such that  $p_{\max}^k < \frac{1}{16}$ . This means that  $1 < k < \lfloor \frac{4}{\log 1/p_{\max}} \rfloor$ . We now carry out the above procedure for the  $k$ -product alphabet. However, since decoding is performed  $k$  symbols at a time, we set our delay threshold to be  $\tilde{d} = \lfloor \frac{d+1}{k} \rfloor - 1$ . Substituting the above into (13) we get

$$\begin{aligned} \mathfrak{R}_n^{\varepsilon'}(P) &\leq 2p_{\max}^{k\tilde{d}} \left(\tilde{d} \log(2/p_{\max}^k) + \kappa + 1\right)^2 \\ &\leq 2p_{\max}^{d-c(p_{\max})} ((d - c(p_{\max})) \log(2/p_{\max}) + \kappa + 1)^2 \end{aligned}$$

**Remark 8.** The scheme described above also allows the encoder to change the delay constraint on the fly, by inserting a suitable fictitious symbol in accordance to the modified constraint. Once the decoder is made aware of this change, both encoder and decoder need to simultaneously adjust the probability of the fictitious symbols.

## B. An Asymptotic Result

**Theorem 3.** The inf-redundancy-delay exponent for a source  $P$  is lower bounded by the Rényi entropy of order 2 of the source, i.e.,

$$\underline{E}(P) \geq H_2(P)$$

*Proof of Theorem 3:* We construct a unit delay encoder for the product source  $P^d$  using fictitious symbols in a similar way as done in Theorem 2, with an additional random coding argument. Let  $<$  be a order of  $\mathcal{X}^d$  such that all super-symbols in the same type class are adjacent (and otherwise arbitrary). Let  $<_{y^d}$  be a new order which is obtained by a rotation of the order  $<$ , making  $y^d$  the smallest element, i.e., the unique order that respects  $<$  for each of the sets  $\{y^d\} \cup \{z^d : y^d < z^d\}$  and  $\{z^d : z^d < y^d\}$ , and where the maximal element in the latter set is the maximal element under  $<_{y^d}$ . Finally, let  $<_{y^d}^+$  be the order of  $\mathcal{X}^{d+} \stackrel{\text{def}}{=} \mathcal{X}^d \cup \{x_L, x_R\}$  that respects

$<_{y^d}$  over  $\mathcal{X}^d$ , such that the arithmetic encoder  $\mathcal{E}$  w.r.t.  $<_{y^d}^+$  matched to  $P_\varepsilon^{d+}$  has the fictitious symbols  $x_L, x_R$  mapped into intervals contained in  $\mathcal{I}^\varepsilon(x^n)_L$  and  $\mathcal{I}^\varepsilon(x^n)_R$ , respectively, and are (say) of the minimal order satisfying this.

Let us now draw an i.i.d. sequence  $(Y_1^d, Y_2^d, \dots)$  with a marginal  $P^d$ , independent of the source sequence. At time instance  $k$  (where time is now w.r.t. the product source), we use an arithmetic encoder w.r.t. the random order  $<_{Y_k^d}$ , and matched to  $P_\varepsilon^d$ . Denote the associated random interval-mapping encoder by  $\mathcal{E}$ . It is easy to see that for any point  $a \in \mathcal{I}^\varepsilon(x^{nd})$ , the probability that the interval corresponding to a type  $Q$  will include  $a$  is upper bounded  $p_{\max}^d$  plus the probability of the type class  $T_Q$  under  $P^d$ , where by Lemma 2 the latter is upper bounded by  $2^{-dD(Q\|P)}$ . By the same Lemma, the probability of any super-symbol within the type class  $T_Q$  is  $2^{-d(D(Q\|P)+H(Q))}$ . Thus,

$$\begin{aligned} \mathbb{P}(a \in \mathcal{I}^\varepsilon(X^{n(d+1)}) | X^{nd} = x^{nd}) \\ \leq \sum_{Q \in \mathcal{P}^d(\mathcal{X})} (2^{-dD(Q\|P)} + p_{\max}^d) 2^{-d(D(Q\|P)+H(Q))} \end{aligned} \quad (14)$$

Taking the limit as  $d \rightarrow \infty$ , and since there is only a polynomial number of types, we obtain

$$\begin{aligned} \lim_{d \rightarrow \infty} -\frac{1}{d} \log \mathbb{P}(a \in \mathcal{I}^\varepsilon(X^{n(d+1)}) | X^{nd} = x^{nd}) \\ \geq \inf_{Q \in \mathcal{P}(\mathcal{X})} \left\{ D(Q\|P) + H(Q) + \min \left( D(Q\|P), \log \frac{1}{p_{\max}} \right) \right\} \end{aligned}$$

Let  $V(Q)$  denote the function over which the infimum above is taken, and assume without loss of generality that  $P$  is strictly nonzero over  $\mathcal{X}$ .  $V(Q)$  is continuous and the infimum is taken over a compact set, hence is attained for some  $Q^* \in \mathcal{P}(\mathcal{X})$ . Suppose that  $D(Q^*\|P) > \log 1/p_{\max}$ . Let  $x \in \mathcal{X}$  be such that  $P(x) = p_{\max}$ , and suppose there exists  $y \in \mathcal{X}$  such that  $P(y) < p_{\max}$  and  $Q^*(y) > 0$ . Generate a perturbed distribution  $Q^\dagger$  by increasing the probability assigned by  $Q^*$  to  $x$  by some  $\beta > 0$ , and decreasing the probability assigned by  $Q^*$  to  $y$  by the same  $\beta$ , leaving the other probabilities unchanged. This implies that

$$D(Q^\dagger\|P) + H(Q^\dagger) < D(Q^*\|P) + H(Q^*),$$

since the above is equivalent (by direct calculation) to  $\beta \log(P(x)/P(y)) > 0$ , which holds true under the assumptions made. Now, by continuity, there exists  $\beta$  small enough such that  $D(Q^\dagger\|P) > \log 1/p_{\max}$ . Hence  $V(Q^\dagger) < V(Q^*)$  for such  $\beta$ , contradicting the minimality of  $Q^*$ . If such  $y$  does not exist, then  $P(x) = p_{\max}$  over the entire support of  $Q^*$ . Therefore,  $D(Q^*\|P) = \log 1/p_{\max} - H(Q^*) \leq \log 1/p_{\max}$ , in contradiction to our assumption. We conclude that  $D(Q^*\|P) \leq \log 1/p_{\max}$ . Hence,

$$\begin{aligned} \lim_{d \rightarrow \infty} -\frac{1}{d} \log \mathbb{P}(a \in \mathcal{I}^\varepsilon(X^{n(d+1)}) | X^{nd} = x^{nd}) \\ = \min_{Q \in \mathcal{P}(\mathcal{X})} \{2D(Q\|P) + H(Q)\} = H_2(P) \end{aligned}$$

where Lemma 3 was invoked in the last equality. Continuing this line of argument, we can essentially replace  $p_{\max}^d$  with

$2^{-dH_2(P)}$  for  $d$  large enough, throughout our proofs. Therefore, the redundancy averaged over the ensemble of random  $d$ -delay constrained encoders is bounded by

$$\mathbb{E}(\mathfrak{R}^{\mathcal{E}}(P)) = O(2^{-dH_2(P)}) \quad (15)$$

and thus there exists a deterministic encoder  $\mathcal{E}$  achieving at least that expected performance, concluding the proof. ■

## V. AN UPPER BOUND FOR $\overline{E}(P)$

In this section we prove an upper bound on the sup-redundancy–delay exponent, hence obtaining an asymptotic lower bound for the redundancy–delay function. This characterizes the best possible redundancy achievable by any delay-constrained encoder. Our bound holds for *almost any* memoryless source, which is meant w.r.t. the Lebesgue measure over the probability simplex.

**Theorem 4.** *For almost any memoryless source  $P$ , the sup-redundancy–delay exponent is upper bounded by*

$$\overline{E}(P) \leq 8 \log \left( \frac{|\mathcal{X}|}{p_{\min}} \right) \quad (16)$$

**Remark 9.** *Note that (16) cannot hold for all sources, e.g. for 2-adic sources we can have zero redundancy with zero delay, hence an infinite exponent.*

**Remark 10.** *When restricted to interval–mapping encoders only, a tighter upper bound of  $8 \log(1/p_{\min})$  holds.*

### A. Proof Outline

Since the proof is somewhat tedious, we find it instructive to provide a rough outline under the assumption that the encoder admits an interval–mapping representation (rather than a generalized one). This assumption will be removed in the proof itself. Due to the strict delay constraint, at any time instance the encoder must map the next  $d$  symbols into intervals that do not contain any forbidden points<sup>11</sup>. Typically (for almost every interval), we will find an infinite number of forbidden points concentrated near the edges, with a typical “concentration region” whose size depends on the specific interval. Clearly, the distances between consecutive points diminishes exponentially to zero. Therefore, mapping symbols to the concentration region will result in a significant mismatch between the symbol probability and the interval length, and this phenomena incurs redundancy. This observation is made precise in Lemma 14.

Now, loosely speaking, there are two opposing strategies the encoder may use when mapping symbols to intervals. The first is to think short-term, namely to be as faithful to the source as possible by assigning interval lengths closely matching symbol probabilities (within the forbidden points constraint). This will likely cause the next source interval to have a relatively large concentration region, resulting in an inevitable redundancy at the subsequent mapping. The second strategy is to think long-term, by mapping to intervals with a small concentration

region. This in general cannot be done while still being faithful to the source’s distribution, hence this strategy also incurs in an inevitable redundancy. The latter observation is made precise in Lemma 18. Our bound results from the tension between these two counterbalancing sources of redundancy.

### B. Proof of Theorem 4

In light of Lemma 6, we can restrict our discussion to generalized interval–mapping encoders of the form (3). However, we will find it more convenient to consider a broader family of generalized interval–mapping encoders, satisfying the following conditions:

- (i) For any  $s \in \mathcal{X}^*$ ,  $\mathcal{I}^{\mathcal{E}}(s)$  is a union of at most  $|\mathcal{X}|^d$  intervals.<sup>12</sup>
- (ii) For any  $s \in \mathcal{X}^*$ ,  $x^d \in \mathcal{X}^d$ ,  $\mathcal{I}^{\mathcal{E}}(sx^d)$  contains no forbidden points from any of the intervals comprising  $\mathcal{I}^{\mathcal{E}}(s)$ .<sup>13</sup>

Let  $I \subseteq [0, 1)$  be a finite union of disjoint intervals  $\{I_k\}_{k=1}^K$ . Recall that  $S_0(I_k)$  is the set of all forbidden points in the interval  $I_k$ . Define:

$$A(I) \stackrel{\text{def}}{=} \bigcup_{k=1}^K \left\{ \frac{|a-b|}{|I|} : a, b \in S_0(I_k), (a, b) \cap S_0(I_k) = \emptyset \right\}$$

and let

$$\delta_I = \delta_I(P, d) \stackrel{\text{def}}{=} \max\{a \in A(I) : a < p_{\min}^d/4\}$$

Namely,  $\delta_I$  is the maximal distance between two consecutive forbidden points in some  $I_k$ , normalized by the measure of  $I$ , that is smaller than  $p_{\min}^d/4$ .

**Lemma 14.**  $r_d(x^n) > \delta_{\mathcal{I}^{\mathcal{E}}(x^n)}$

*Proof:* See the Appendix. ■

A number  $a \in [0, 1)$  is called  $(m, \ell)$ –constrained if

$$a = 0.\underbrace{00\dots 0}_{m'(a)}\underbrace{1\phi\dots\phi}_m\underbrace{00\dots 0}_\ell\phi\dots$$

where  $m'(a)$  is the length of the zeros prefix of  $a$ , and  $\phi$  is the “don’t care” symbol. The  $(m, \ell)$ –constrained region  $\mathcal{C}_{m,\ell}$  is the set of all such numbers. A number  $a \in [0, 1)$  is called  $(m, \ell)$ –violating if

$$a = 0.\underbrace{00\dots 0}_{m'(a)}\underbrace{1\phi\dots\phi}_m\underbrace{\phi\dots\dots\dots\phi}_{\ell \text{ bits, not all '0' or all '1'}}\phi\dots \quad (17)$$

The  $(m, \ell)$ –violating region  $\mathcal{V}_{m,\ell}$  is the set of all such numbers. The complement  $\overline{\mathcal{V}}_{m,\ell} = [0, 1) \setminus \mathcal{V}_{m,\ell}$  is called the  $(m, \ell)$ –permissible region. Define the regions<sup>14</sup>

$$LC_{m,\ell} \stackrel{\text{def}}{=} \langle -\log \mathcal{C}_{m,\ell} \rangle, \quad L\overline{\mathcal{V}}_{m,\ell} \stackrel{\text{def}}{=} \langle -\log \overline{\mathcal{V}}_{m,\ell} \rangle$$

and let

$$\mathcal{D}_{m,\ell}^{(1)} \stackrel{\text{def}}{=} \langle L\overline{\mathcal{V}}_{m,\ell} - LC_{m,\ell} \rangle, \quad \mathcal{D}_{m,\ell}^{(2)} \stackrel{\text{def}}{=} \langle \mathcal{D}_{m,\ell}^{(1)} - \mathcal{D}_{m,\ell}^{(1)} \rangle$$

<sup>12</sup>To disambiguate the statement, we clarify that any two intervals whose union is an interval are counted as a single interval.

<sup>13</sup>Note that this is satisfied by (3), since  $\text{bin}(\mathcal{I}^{\mathcal{E}}(sx^d))$  is always contained in one of the intervals comprising  $\mathcal{I}^{\mathcal{E}}(s)$ .

<sup>14</sup>The log and  $\langle \cdot \rangle$  operations are taken pointwise on the set elements.

<sup>11</sup>As mentioned in Remark 7, avoiding forbidden points is not always a necessary condition. However, in the next section we verify this is not a restriction.

The following two lemmas are easily observed.

**Lemma 15.** Let  $\mu > 0$ . If  $a \in \mathcal{V}_{m,\ell}$  and  $b \in \mathcal{C}_{m,\ell'}$  where  $\ell < \ell'$ , then

$$|a - b| \geq 2^{-m'(a)} \cdot 2^{-(m+\ell)} \geq \frac{a}{2} \cdot 2^{-(m+\ell)}$$

**Lemma 16.** If  $I, J \subseteq [0, 1)$  are each a union of at most  $M$  intervals of size no larger than  $r$  each, then  $\langle I - J \rangle$  can be written as a union of at most  $M^2 + 1$  intervals of size no larger than  $2r$  each.

The  $(m, \ell)$ -permissible region within the interval  $[1/2, 1)$  is comprised of  $2^{m-1} + 1$  subintervals. By definition, the size of each is upper-bounded by  $2^{-(m'+m+\ell)+1}$ . Applying  $\langle -\log(\cdot) \rangle$  to all such intervals in the  $[1/2, 1)$  interval (corresponding to  $m' = 0$ ) will stretch each of them by a factor of at most  $2 \log e < 4$ . All other permissible intervals (those with  $m' > 0$ ) coincide on the unit interval after applying the  $\langle -\log(\cdot) \rangle$  operator. Hence  $L\bar{\mathcal{V}}_{m,\ell}$  can be written as a union of at most  $2^{m-1} + 1$  intervals, each of size at most  $2^{-(m+\ell)+3}$ . A similar argument shows that  $L\bar{\mathcal{V}}_{m,\ell}^{(1)}$  can also be written that way<sup>15</sup>. Appealing to Lemma 16,  $\mathcal{D}_{m,\ell}^{(1)}$  can be written as a union of at most  $(2^{m-1} + 1)^2 + 1$  intervals, each of size at most  $2^{-(m+\ell)+4}$ . Applying the Lemma again, we find that  $\mathcal{D}_{m,\ell}^{(2)}$  can be written as a union of at most  $((2^{m-1} + 1)^2 + 1)^2 + 1 \leq 2^{4m+1}$  intervals each of size at most  $2^{-(m+\ell)+5}$ . Hence,

$$|\mathcal{D}_{m,\ell}^{(2)}| < 2^{4m+1} \cdot 2^{-(m+\ell)+5} = 2^{3m-\ell+6} \quad (18)$$

A source  $P$  is called  $(\mu_0, \lambda)$ -regular if there exists a pair of symbols  $y, z \in \mathcal{X}$  and  $m_0 \in \mathbb{N}$  such that for any  $\mu \geq \mu_0$

$$\lambda = \left\langle \log \frac{P(y)}{P(z)} \right\rangle \notin \bigcup_{m=m_0}^{\infty} \mathcal{D}_{m, \lceil \mu m \rceil}^{(2)} \quad (19)$$

**Remark 11.**  $0 \in \mathcal{D}_{m, \lceil \mu m \rceil}^{(2)}$  for any  $m$  and  $\mu$ , hence no source can be  $(\mu_0, 0)$ -regular. Since for a dyadic source  $\lambda = 0$  for any pair  $y, z$ , a dyadic source is never  $(\mu_0, \lambda)$ -regular.

The following two lemmas establish some properties of  $(\mu_0, \lambda)$ -regularity.

**Lemma 17.** Let  $\mu_0 > 3$ . Almost any source is  $(\mu_0, \lambda)$ -regular for some  $\lambda > 0$ .

*Proof:* See the Appendix. ■

Define the following set:

$$A_{\alpha,\beta}^d \stackrel{\text{def}}{=} \left\{ x^d \in \mathcal{X}^d : \langle -\log P(x^d) \rangle \notin \mathcal{D}_{\lceil \alpha d \rceil, \lceil \beta d \rceil}^{(1)} \right\}$$

**Lemma 18.** Suppose  $P$  is a  $(\mu_0, \lambda)$ -regular source. Then for any  $\alpha, \beta > 0$  with  $\beta/\alpha > \mu_0$

$$\liminf_{d \rightarrow \infty} P(A_{\alpha,\beta}^d) \geq \frac{1}{2}$$

*Proof:* See the Appendix. ■

<sup>15</sup>It can in fact be written as a union of less and smaller intervals, but that adds nothing to our argument.

From this point forward we assume  $P$  is  $(\mu_0, \lambda)$ -regular with  $\mu_0 > 3$ . Let  $\mu < \mu'$ , and define the indexed sets

$$B_k \stackrel{\text{def}}{=} \left\{ x^k \in \mathcal{X}^k : \delta_{\mathcal{I}^{\mathcal{E}}(x^k)} > p_{\min}^{\mu d} \right\}$$

$$C(x^k) \stackrel{\text{def}}{=} \left\{ y^d \in \mathcal{X}^d : \delta_{\mathcal{I}^{\mathcal{E}}(x^k y^d)} > p_{\min}^{\mu' d} \right\}$$

For  $x^k \in B_k$ , Lemma 14 implies that

$$r_d(x^k) > p_{\min}^{\mu d} \quad (20)$$

On the other hand,  $x^k \notin B_k$  implies that the length of each interval comprising  $\mathcal{I}^{\mathcal{E}}(x^k)$  must be in  $\mathcal{C}_{\lceil d \log(1/p_{\min}) \rceil, \lceil \mu d \log(1/p_{\min}) \rceil}$ . Since there are at most  $|\mathcal{X}|^d$  such intervals, it must be that

$$|\mathcal{I}^{\mathcal{E}}(x^k)| \in \mathcal{C}_{\lceil \alpha d \rceil, \lceil \beta d \rceil} \quad (21)$$

where

$$\alpha \stackrel{\text{def}}{=} \log(1/p_{\min}) + \log |\mathcal{X}|, \quad \beta \stackrel{\text{def}}{=} \mu \log(1/p_{\min}) - \log |\mathcal{X}|$$

Similarly, if  $y^d \notin C(x^k)$  then

$$|\mathcal{I}^{\mathcal{E}}(x^k y^d)| \in \mathcal{C}_{\lceil \alpha d \rceil, \lceil \beta' d \rceil} \quad (22)$$

where

$$\beta' \stackrel{\text{def}}{=} \mu' \log(1/p_{\min}) - \log |\mathcal{X}|$$

For Lemma 18 to apply, we set  $\mu, \mu'$  such that  $\beta/\alpha > \mu_0$  and  $\beta'/\alpha > \mu_0$ . This yields the constraints:

$$\mu' > \mu > \mu_0 + \frac{(\mu_0 + 1) \log |\mathcal{X}|}{\log(1/p_{\min})}$$

In what follows, we will think of  $\mu'$  as arbitrarily close to  $\mu$ . For any  $x^k \notin B_k$  we have:

$$\begin{aligned} & \mathbb{E}(r_d(X^k) + r_d(X^{k+d}) \mid X^k = x^k) \\ & \stackrel{(a)}{\geq} \left( \sum_{y^d \in A_{\alpha,\beta}^d \cap \overline{C(x^k)}} |P(y^d) - \mu_d^{\mathcal{E}}(y^d | x^k)| \right)^2 \\ & \quad + p_{\min}^{\mu' d} P(C(x^k)) \\ & = \left( \sum_{y^d \in A_{\alpha,\beta}^d \cap \overline{C(x^k)}} \left| \frac{P(y^d) |\mathcal{I}^{\mathcal{E}}(x^k)| - |\mathcal{I}^{\mathcal{E}}(x^k y^d)|}{|\mathcal{I}^{\mathcal{E}}(x^k)|} \right| \right)^2 \\ & \quad + p_{\min}^{\mu' d} P(C(x^k)) \\ & \stackrel{(b)}{\geq} \left( \frac{1}{|\mathcal{I}^{\mathcal{E}}(x^k)|} \sum_{y^d \in A_{\alpha,\beta}^d \cap \overline{C(x^k)}} \frac{P(y^d) |\mathcal{I}^{\mathcal{E}}(x^k)|}{2} p_{\min}^{\lceil \alpha d \rceil + \lceil \beta d \rceil} \right)^2 \\ & \quad + p_{\min}^{\mu' d} P(C(x^k)) \\ & = \left( \frac{P(A_{\alpha,\beta}^d \cap \overline{C(x^k)})}{2} \right)^2 \cdot p_{\min}^{2(\alpha+\beta)d+4} + p_{\min}^{\mu' d} P(C(x^k)) \\ & \stackrel{(c)}{\geq} \frac{1}{4} \left[ (P(A_{\alpha,\beta}^d \cap \overline{C(x^k)})^2 + P(C(x^k))) \right] p_{\min}^{d \max(2(\alpha+\beta), \mu') + 4} \\ & \stackrel{(d)}{\geq} \frac{1}{4} \left[ (P(A_{\alpha,\beta}^d) - P(A_{\alpha,\beta}^d \cap C(x^k)))^2 + P(A_{\alpha,\beta}^d \cap C(x^k)) \right] \\ & \quad \times p_{\min}^{2d(\mu+1) \log(1/p_{\min}) + 4} \end{aligned}$$

$$\begin{aligned}
&\stackrel{(e)}{\geq} \frac{1}{4} (P(A_{\alpha,\beta}^d))^2 \cdot p_{\min}^{2d(\mu+1)\log(1/p_{\min})+4} \\
&= \left(\frac{1}{16} + o(1)\right) \cdot p_{\min}^{2d(\mu+1)\log(1/p_{\min})+4} \quad (23)
\end{aligned}$$

The inequalities are justified as follows:

- (a) Pinsker's inequality for the divergence [1] was used, together with Lemma 14 and the nonnegativity of  $r_d(\cdot)$ .
- (b) (21) and (22) hold for all the union-of-intervals lengths in the summation. Since  $\langle -\log P(y^d) \rangle \notin \mathcal{D}_{[\alpha d], [\beta d]}^{(1)}$  for each  $y^d$  in the summation, then appealing to Lemma 1, we have that  $P(y^d) |Z^{\mathcal{E}}(x^k)| \in \mathcal{V}_{[\alpha d], [\beta' d]}$ . The inequality now follows by virtue of Lemma 15.
- (c)  $P(A \cap \overline{C}) = P(A) - P(A \cap C)$  and  $P(C) \geq P(A \cap C)$ .
- (d)  $\mu'$  can be taken to be arbitrarily close to  $\mu$ .
- (e) Lemma 18 was used to lower bound the probability of the set  $A_{\alpha,\beta}^d$ .

Combining (20) and (23), we get:

$$\begin{aligned}
&\mathbb{E}(r_d(X^k) + r_d(X^{k+d})) \\
&\geq \min \left( p_{\min}^{\mu d}, \left( \frac{1}{16} + o(1) \right) \cdot p_{\min}^{2d(\mu+1)\log(1/p_{\min})+4} \right) \\
&= \left( \frac{1}{16} + o(1) \right) \cdot p_{\min}^{2d(\mu+1)\log(1/p_{\min})+4}
\end{aligned}$$

This holds for any  $d$ -constrained encoder  $\mathcal{E} \in \mathcal{C}_d$ , hence and plugging into Lemma 7 we get

$$\begin{aligned}
\underline{\mathcal{R}}^{\mathcal{E}}(P) &= \liminf_{n \rightarrow \infty} \frac{1}{2nd} \sum_{k=1}^n \mathbb{E}(r_d(X^k) + r_d(X^{k+d})) \\
&\geq \left( \frac{1}{16} + o(1) \right) \cdot \frac{1}{2d} \cdot p_{\min}^{2d(\mu+1)\log(1/p_{\min})+4}
\end{aligned}$$

This lower bound holds for any  $\mu > \mu_0 + \frac{(\mu_0+1)\log|\mathcal{X}|}{\log(1/p_{\min})}$ . Moreover, by Lemma 17 almost any source is  $(\mu_0, \lambda)$ -regular for any  $\mu_0 > 3$ . Therefore, we have that for almost any source

$$\underline{\mathcal{R}}^{\mathcal{E}}(P) \geq \left( \frac{1}{16} + o(1) \right) \cdot \frac{1}{2d} \cdot p_{\min}^{8d\log(\frac{|\mathcal{X}|}{p_{\min}}) + o(d)}$$

and hence

$$\overline{E}(P) \leq 8 \log \left( \frac{|\mathcal{X}|}{p_{\min}} \right)$$

As mentioned in Remark 10, if the encoder is restricted to be interval-mapping then a tighter upper bound  $8 \log(1/p_{\min})$  holds. In this case  $\mathcal{I}^{\mathcal{E}}(\cdot)$  is a single interval rather than a union of  $|\mathcal{X}|^d$  intervals, hence the proof remains the same up to the substitution  $|\mathcal{X}| \leftrightarrow 1$ .

## VI. CONCLUSIONS

The redundancy in lossless coding of a memoryless source incurred by imposing a strict end-to-end delay constraint was analyzed, and shown to decay exponentially with the delay. The associated delay-redundancy exponent was lower bounded by the Rényi entropy  $H_2(P)$  for any source  $P$ , and upper bounded by  $8 \log(|\mathcal{X}|/p_{\min})$  for most sources. This exponential behavior should be juxtaposed against classical results in source coding, showing a polynomial decay of the redundancy with the delay. In the classical framework, the delay is identified with the block length or the maximal phrase

length, which in our framework imposes a harsh restriction: The decoder is not allowed to start reproducing source symbols in the midst of a block/phrase, and the delay is repeatedly nullified at the end of each block/phrase. This means the encoder is reset at these instances, i.e., the prefix has no effect on its future behavior. Loosely speaking, the gain of exponential versus polynomial is reaped via a tighter control over the delay process, making such reset events rare. This superior performance comes however at a possible cost: in contrast to the block/phrase-constrained setup where the encoder can clear its memory and start-over in roughly constant intervals, the more general encoders discussed in this paper need to keep track of a state. The precision required for keeping the state is however finite, and can be easily derived from Lemma 14.

In our framework, we have isolated the impact of the delay on the redundancy by letting the transmission time  $n$  go to infinity. This also makes sense complexity-wise, since the per-symbol encoding complexity is determined primarily by the delay, and not by the length of the encoded sequence. In practice however, a finite transmission time forces the encoder to terminate the codeword, which in turn incurs an additional penalty of  $O(n^{-1})$  in redundancy. Setting  $d = O(\log n)$  renders this additional redundancy term commensurate with the redundancy incurred by the delay constraint. Therefore, our results imply that the delay can be made logarithmic in the block length, while maintaining the same order of redundancy. Conversely, for almost all sources this is the best possible tradeoff between block length and delay. A similar statement in the context of universal source coding was mentioned in [27], though for a somewhat different definition of the delay.

There is still a large gap between the lower and upper bounds on the redundancy-delay exponent, where the upper bound seems particularly loose. Furthermore, it remains to be seen whether the zero-measure set of sources for which the upper bound may fail to hold, can be reduced from the set of sources that do not satisfy our intricate regularity condition, to the set of dyadic sources only, which is the smallest possible.

## APPENDIX

*Proof of Lemma 6:* Let us first show that  $\mathcal{I}^{\mathcal{E}}$  satisfies the conditions for a generalized interval-mapping encoder.  $\mathcal{I}^{\mathcal{E}}(sx) \subseteq \mathcal{I}^{\mathcal{E}}(s)$  is immediate from the consistency property. Let  $y, z \in \mathcal{X}$  be distinct, and assume that  $\mathcal{I}^{\mathcal{E}}(sy) \cap \mathcal{I}^{\mathcal{E}}(sz) \neq \emptyset$ . Then since any two binary intervals are either disjoint or one is contained in the other, then without loss of generality there exist  $x^d, \tilde{x}^d$  such that  $[\mathcal{E}(syx^d)] \subseteq [\mathcal{E}(sz\tilde{x}^d)]$ , i.e., such that  $\mathcal{E}(sz\tilde{x}^d) \preceq \mathcal{E}(syx^d)$ . Since  $\delta^{\mathcal{E}}(\cdot, \cdot) \leq d$ , it must be that  $sz \preceq syx^d$ , in contradiction. This verifies the disjoint nesting property.

By the consistency property,  $\mathcal{I}^{\mathcal{E}}(s) \subseteq [\mathcal{E}(s)]$ . Suppose that there exists a binary interval  $[b]$  such that  $\mathcal{I}^{\mathcal{E}}(s) \subseteq [b] \subset [\mathcal{E}(s)]$ . Then  $\mathcal{E}(s) \prec b \preceq \mathcal{E}(sx^d)$  for any  $x^d \in \mathcal{X}^d$ , and hence by the integrity property it must be that  $b \preceq \mathcal{E}(s)$ , in contradiction. Hence  $\text{bin}(\mathcal{I}^{\mathcal{E}}(s)) = [\mathcal{E}(s)]$  for any  $s \in \mathcal{X}^*$ , verifying the minimality property. ■

*Proof of Lemma 7:*

(i)

$$\begin{aligned}
\mathfrak{R}_n^\mathcal{E}(P) &= \bar{L}_n^\mathcal{E} - H(P) \\
&= \frac{1}{n} \mathbb{E}(-\log |\text{bin}(\mathcal{I}^\mathcal{E}(X^n))|) - H(P) \\
&\leq \frac{1}{n} (\mathbb{E}(-\log \mu^\mathcal{E}(X^n)) - H(P^n)) \\
&= \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P(x^n) \log \left( \frac{P(x^n)}{\mu^\mathcal{E}(x^n)} \right) \\
&= R_n^\mathcal{E}(P)
\end{aligned}$$

(ii) Consider the generalized interval mapping representation of  $\mathcal{E}$  given in Lemma 6. This representation satisfies  $\mathcal{I}^\mathcal{E}(x^{n+d}) \subseteq \mathcal{I}^\mathcal{E}(x^n)$ . Thus similarly to the above:

$$\begin{aligned}
\mathfrak{R}_n^\mathcal{E}(P) &= \frac{1}{n} \mathbb{E}(-\log |\text{bin}(\mathcal{I}^\mathcal{E}(X^n))|) - H(P) \\
&\geq \frac{1}{n} \left( \mathbb{E}(-\log \mu^\mathcal{E}(X^{n+d})) - \frac{n}{n+d} H(P^{n+d}) \right) \\
&= \left( \frac{n+d}{n} \right) R_{n+d}^\mathcal{E}(P) + \frac{d}{n} H(P)
\end{aligned}$$

(iii) For any fixed  $d \in \mathbb{N}$ ,

$$\begin{aligned}
&\frac{1}{nd} \sum_{k=1}^n \mathbb{E} r_d(X^k) \\
&= -H(P) + \mathbb{E} \left( \frac{1}{nd} \sum_{k=1}^n \log \frac{\mu_k^\mathcal{E}(X^k)}{\mu_{k+d}^\mathcal{E}(X^{k+d})} \right) \\
&= -H(P) + \frac{1}{nd} \sum_{k=1}^d \mathbb{E} \log \mu_k^\mathcal{E}(X^k) \\
&\quad - \frac{1}{nd} \sum_{k=1}^d \mathbb{E} \log \mu_{n+k}^\mathcal{E}(X^{n+k}) \\
&\leq O(n^{-1}) - H(P) - \frac{1}{n} \mathbb{E} \log \mu_{n+d}^\mathcal{E}(X^{n+d}) \\
&= O(n^{-1}) + \left( \frac{n+d}{n} \right) R_{n+d}^\mathcal{E} + \frac{d}{n} H(P) \\
&\leq \mathfrak{R}_n^\mathcal{E} + O(n^{-1})
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{1}{nd} \sum_{k=1}^n \mathbb{E} r_d(X^k) &\geq O(n^{-1}) - H(P) - \frac{1}{n} \mathbb{E} \log \mu_n^\mathcal{E}(X^n) \\
&= R_n^\mathcal{E} + O(n^{-1}) \geq \mathfrak{R}_n^\mathcal{E} + O(n^{-1})
\end{aligned}$$

*Proof of Lemma 8:* We only need to prove (i). An arithmetic encoder matched to the source  $P$  is well known to achieve zero asymptotic redundancy [11], and a bounded expected delay [14], [15], [16]. Therefore

$$\inf_{\mathcal{E} \in \mathfrak{L}(P)} \bar{\mathfrak{R}}^\mathcal{E}(P) \leq \inf_{\mathcal{E} \in \mathfrak{B}(P)} \bar{\mathfrak{R}}^\mathcal{E}(P) \leq 0$$

Let  $\mathcal{E} \in \mathfrak{L}(P)$ . Define  $B_d$  to be the set of all suffixes that allow decoding of any prefix with delay at most  $d$ , i.e.,

$$B_d \stackrel{\text{def}}{=} \{y^\infty \in \mathcal{X}^\infty : \delta^\mathcal{E}(s, y^\infty) \leq d, \forall s \in \mathcal{X}^*\}$$

The lossless property implies that for any  $\varepsilon > 0$  there exists  $d$  large enough such that

$$P(B_d) \geq 1 - \varepsilon \quad (24)$$

Define  $\bar{B}_d$  to be the set of all prefixes in  $B_d$ , i.e.,

$$\bar{B}_d \stackrel{\text{def}}{=} \{z^d \in \mathcal{X}^d : z^d \prec y^\infty \in B_d\}$$

Note that by the very definition of  $B_d$ , each prefix in  $\bar{B}_d$  must appear in  $B_d$  with all possible suffixes. Therefore,  $P(\bar{B}_d) = P(B_d) \geq 1 - \varepsilon$  for  $d$  large enough. Furthermore the lossless property also implies that for any  $z^d \in \bar{B}_d$ , the BV codebook  $C_{z^d} : \mathcal{X}^n \mapsto \{0, 1\}^*$  defined by

$$C_{z^d}(x^n) \stackrel{\text{def}}{=} \mathcal{E}(x^n z^d) \quad (25)$$

is a prefix-free lossless codebook, and hence must satisfy  $\mathbb{E}|C_{z^d}(X^n)| \geq nH(P)$ . Write:

$$\begin{aligned}
\bar{L}_{n+d}^\mathcal{E}(P) &= \frac{1}{n+d} \sum_{z^d \in \mathcal{X}^d} P(z^d) \sum_{x^n \in \mathcal{X}^n} P(x^n) |\mathcal{E}(x^n z^d)| \\
&\geq \frac{1}{n+d} \sum_{z^d \in \bar{B}_d} P(z^d) \sum_{x^n \in \mathcal{X}^n} P(x^n) |\mathcal{E}(x^n z^d)| \\
&\geq \frac{1}{n+d} \sum_{z^d \in \bar{B}_d} P(z^d) \mathbb{E}|C_{z^d}(X^n)| \\
&\geq \frac{1}{n+d} \cdot P(\bar{B}_d) \cdot nH(P) \geq \frac{(1-\varepsilon)n}{n+d} H(P)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\underline{\mathfrak{R}}^\mathcal{E} &= \liminf_{n \rightarrow \infty} \mathfrak{R}_{n+d}^\mathcal{E}(P) \geq \lim_{n \rightarrow \infty} \left( \frac{(1-\varepsilon)n}{n+d} - 1 \right) H(P) \\
&= -\varepsilon H(P)
\end{aligned}$$

This holds for any  $\varepsilon > 0$ , hence  $\underline{\mathfrak{R}}^\mathcal{E} \geq 0$ .  $\blacksquare$

*Proof of Lemma 9:* Let  $\mathcal{E} \in \mathfrak{C}_d$ , and set any  $\varepsilon > 0$ . We show that there exists another encoder  $\mathcal{E}' \in \mathfrak{C}_d$  such that

$$\bar{\mathfrak{R}}^{\mathcal{E}'}(P) \leq \underline{\mathfrak{R}}^\mathcal{E}(P) + \varepsilon$$

which immediately establishes the Lemma. The encoder  $\mathcal{E}'$  will be constructed by properly terminating  $\mathcal{E}$ . Set  $n$  large enough such that both

$$n > d + \min\left\{d, \frac{2d\underline{\mathfrak{R}}^\mathcal{E}(P)}{\varepsilon}\right\} \quad (26)$$

and

$$\mathfrak{R}_n^\mathcal{E}(P) \leq \underline{\mathfrak{R}}^\mathcal{E}(P) + \varepsilon/4 \quad (27)$$

$\blacksquare$  For any  $x^{n-d} \in \mathcal{X}^{n-d}$ , define

$$y^d(x^{n-d}) \stackrel{\text{def}}{=} \underset{z^d \in \mathcal{X}^d}{\text{argmin}} \{|\mathcal{E}(x^{n-d} z^d)|\}$$

namely,  $y^d(x^{n-d})$  is the suffix that results in the minimal codelength after having encoded  $x^{n-d}$ . Clearly,

$$n^{-1} \mathbb{E}|\mathcal{E}(X^{n-d} y^d(X^{n-d}))| \leq \bar{L}_n^\mathcal{E}(P) \quad (28)$$

Construct the new encoder  $\mathcal{E}'$  as follows. For any  $k < n-d$ , let  $\mathcal{E}'(x^k) = \mathcal{E}(x^k)$ , and let  $\mathcal{E}'(x^{n-d}) = \mathcal{E}(x^{n-d} y^d(x^{n-d}))$ . For  $k > n-d$ , divide  $x^k$  into blocks of equal size  $n-d$  (with the last one possibly shorter), apply the rule above to each

separately, and let  $\mathcal{E}'(x^k)$  be the concatenation thereof. Using (28), we have

$$\begin{aligned}\mathfrak{R}_{n-d}^{\mathcal{E}'}(P) &= (n-d)^{-1} \mathbb{E}|\mathcal{E}'(X^{n-d})| - H(P) \\ &\stackrel{(a)}{\leq} \frac{n}{n-d} \bar{L}_n^{\mathcal{E}}(P) - H(P) \leq \frac{n}{n-d} \mathfrak{R}_n^{\mathcal{E}}(P) \\ &\stackrel{(b)}{\leq} \underline{\mathfrak{R}}^{\mathcal{E}}(P) + \left( \frac{d}{n-d} \mathfrak{R}^{\mathcal{E}}(P) + \frac{n}{n-d} \cdot \varepsilon/4 \right) \\ &\stackrel{(c)}{\leq} \underline{\mathfrak{R}}^{\mathcal{E}}(P) + \varepsilon\end{aligned}$$

where (a) follows from (28), (b) follows from (27), and (c) follows from the assumption (26). Now, from the concatenated construction we have that for any  $m > n-d$

$$\begin{aligned}\mathfrak{R}_m^{\mathcal{E}'}(P) &\leq \frac{\lceil m/(n-d) \rceil}{m} \cdot (n-d) \cdot \mathfrak{R}_{n-d}^{\mathcal{E}'}(P) \\ &\leq \frac{m+n-d}{m} (\underline{\mathfrak{R}}^{\mathcal{E}}(P) + \varepsilon)\end{aligned}$$

and hence

$$\overline{\mathfrak{R}}^{\mathcal{E}'}(P) = \limsup_{m \rightarrow \infty} \mathfrak{R}_m^{\mathcal{E}'}(P) \leq \underline{\mathfrak{R}}^{\mathcal{E}}(P) + \varepsilon$$

as desired.  $\blacksquare$

*Proof of Lemma 10:* It is easy to see that the number of t-left-adjacents of  $p$  that are larger than  $a + \delta$  is the number of ones in the binary expansion of  $(p-a)$  up to resolution  $\delta$ . Similarly, the number of t-right-adjacents of  $p$  that are smaller than  $b - \delta$  is the number of ones in the binary expansion of  $(b-p)$  up to resolution  $\delta$ . Defining  $\lceil x \rceil^+ \stackrel{\text{def}}{=} \max(\lceil x \rceil, 0)$ , we get:

$$\begin{aligned}|S_\delta(I, p)| &\leq \lceil \log \frac{p-a}{\delta} \rceil^+ + \lceil \log \frac{b-p}{\delta} \rceil^+ \\ &\leq \begin{cases} 2 + \log \frac{(p-a)(b-p)}{\delta^2} & , \delta < p-a, b-p \\ 1 + \log \frac{|b-a|}{\delta} & , o.w. \end{cases} \\ &\leq 1 + 2 \log \frac{|b-a|}{\delta}\end{aligned}$$

*Proof of Lemma 14:* Let  $I = \mathcal{I}^{\mathcal{E}}(x^n)$  throughout the proof. Let

$$z^d \stackrel{\text{def}}{=} \operatorname{argmin}_{y^d \in \mathcal{Y}^d} \mu_d^{\mathcal{E}}(y^d | x^n)$$

and let  $\gamma \stackrel{\text{def}}{=} \mu_d^{\mathcal{E}}(z^d | x^n)$ . If  $\gamma < \delta_I$ , then  $z^d$  has been assigned with a measure at least four times smaller than its probability  $P(z^d)$ . The  $d$ -instantaneous redundancy can be lower bounded as follows:

$$\begin{aligned}r_d(x^n) &= D(P^d \| \mu_d(\cdot | x^n)) \stackrel{(a)}{\geq} D(P(z^d) \| \gamma) \stackrel{(b)}{\geq} D(p_{\min}^d \| \gamma) \\ &= p_{\min}^d \log \frac{p_{\min}^d}{\gamma} + (1 - p_{\min}^d) \log \frac{1 - p_{\min}^d}{1 - \gamma} \\ &\stackrel{(c)}{\geq} 2p_{\min}^d - (1 - p_{\min}^d) \frac{p_{\min}^d}{1 - p_{\min}^d} = p_{\min}^d \geq \delta_I\end{aligned}$$

In (a) we have used the data processing inequality for the divergence<sup>16</sup>. In (b) we have used the fact that  $\gamma < p_{\min}^d \leq$

<sup>16</sup>Recall that  $\mu_d(\cdot | x^n)$  sums to at most unity, hence can be complemented to a probability distribution by adding an auxiliary symbol  $\omega$  to  $\mathcal{X}^d$  and defining  $P^d(\omega) = 0$ .

$P(z^d)$  together with the monotonicity of the scalar relative entropy. In (c) we have used  $\log(1-p) \geq -\frac{p}{1-p}$  for  $0 < p < 1$ .

If on the other hand  $\gamma \geq \delta_I$ , then all of the  $d$ -fold alphabet has been assigned to a measure at most  $1 - \delta_I$  which results in a  $d$ -instantaneous redundancy lower bounded by

$$r_d(x^n) \geq \log \frac{1}{1 - \delta_I} \geq \delta_I \log e \geq \delta_I$$

*Proof of Lemma 17:* Note that  $\mathcal{C}_{m,\ell+1} \subset \mathcal{C}_{m,\ell}$  and  $\mathcal{V}_{m,\ell+1} \supset \mathcal{V}_{m,\ell}$ , hence  $\mathcal{D}_{m,\ell+1}^{(2)} \subset \mathcal{D}_{m,\ell}^{(2)}$ . By (18), we have that for any  $\mu_0 > 3$

$$\begin{aligned}\lim_{m_0 \rightarrow \infty} \left| \bigcup_{\mu \geq \mu_0} \bigcup_{m=m_0}^{\infty} \mathcal{D}_{m,\lceil \mu m \rceil}^{(2)} \right| &= \lim_{m_0 \rightarrow \infty} \left| \bigcup_{m=m_0}^{\infty} \mathcal{D}_{m,\lceil \mu_0 m \rceil}^{(2)} \right| \\ &\leq \lim_{m_0 \rightarrow \infty} \sum_{m=m_0}^{\infty} 2^{m(3-\mu_0)+6} \\ &= \lim_{m_0 \rightarrow \infty} \frac{2^{m_0(3-\mu_0)+6}}{1 - 2^{3-\mu_0}} = 0\end{aligned}$$

The statement of the lemma follows easily.  $\blacksquare$

*Proof of Lemma 18:* We will assume hereinafter that  $\varepsilon < \frac{1}{2}p_{\min}$ . Let  $y, z$  be the symbols attaining  $\lambda$ , and define a transformation  $\sigma : \mathcal{P}^d(\mathcal{X}) \mapsto \mathcal{P}^d(\mathcal{X})$  on types:

$$\sigma(Q)(x) = \begin{cases} Q(x) & x \notin \{y, z\} \quad \vee \quad Q(y) = 0 \\ Q(x) - d^{-1} & x = y \quad \wedge \quad Q(y) > 0 \\ Q(x) + d^{-1} & x = z \quad \wedge \quad Q(y) > 0 \end{cases} \quad (29)$$

Namely,  $\sigma$  exchanges one appearance of  $y$  with the appearance of  $z$  as long as this is possible, i.e., as long as  $Q(y) > 0$ . Now, suppose  $d > \frac{m_0}{\log(1/p_{\min})}$  so that (19) is satisfied. Noting that the set  $A_{\alpha,\beta}^d$  is a union of type classes, let  $Q \in \mathcal{P}_{\varepsilon}^d(\mathcal{X}, P)$  be a type such that  $T_Q \cap A_{\alpha,\beta}^d = \emptyset$ . Clearly  $\sigma(Q) \neq Q$ , and for any  $x^d \in T_Q$  and  $\tilde{x}^d \in T_{\sigma(Q)}$ ,

$$\langle -\log P(\tilde{x}^d) \rangle = \langle -\log P(x^d) + \lambda \rangle$$

Now since  $\lambda \notin \mathcal{D}_{m,\lceil \mu m \rceil}^{(2)}$  for any  $m \geq m_0$  and  $\mu > \mu_0$ , and since  $\beta/\alpha > \mu_0$ , then  $\lambda \notin \mathcal{D}_{\lceil \alpha d \rceil, \lceil \beta d \rceil}^{(2)}$ . Recalling the definition of  $\mathcal{D}_{\lceil \alpha d \rceil, \lceil \beta d \rceil}^{(2)}$  and appealing to Lemma 1, we have that  $\langle -\log P(\tilde{x}^d) \rangle \notin \mathcal{D}_{\lceil \alpha d \rceil, \lceil \beta d \rceil}^{(1)}$ , hence we conclude that  $\sigma(Q) \in A_{\alpha,\beta}^d$ . Therefore, since  $\sigma$  is one-to-one when restricted to  $\mathcal{P}_{\varepsilon}^d(\mathcal{X}, P)$ , then  $\sigma$  uniquely matches any type in  $\mathcal{P}_{\varepsilon}^d(\mathcal{X}, P)$  that is outside  $A_{\alpha,\beta}^d$  to a type that is inside  $A_{\alpha,\beta}^d$ .

Let us now get a handle on the variation in the probability of a type class incurred by applying  $\sigma$ . It is easy to check that for any  $Q \in \mathcal{P}_{\varepsilon}^d(\mathcal{X}, P)$ , and  $n$  large enough,

$$\begin{aligned}P(T_{\sigma(Q)}) &\geq P(T_Q) \left( \frac{(P(y) - \varepsilon)d}{(P(z) + \varepsilon)d + 1} \right) \left( \frac{P(z)}{P(y)} \right) \\ &\geq P(T_Q) \left( 1 - \frac{\varepsilon}{P(y)} \right) \left( 1 - \frac{\varepsilon + d^{-1}}{P(z)} \right) \\ &= P(T_Q) (1 + O(\varepsilon) + O(d^{-1}))\end{aligned}$$

Namely, the probability of a type class for a type  $Q \in \mathcal{P}_{\varepsilon}^d(\mathcal{X}, P)$  under  $P$ , remains almost the same after applying

$\sigma$ . Therefore:

$$\begin{aligned}
& 1 - P(A_{\alpha,\beta}^d) \\
& \leq P\left(\bigcup_{Q \notin \mathcal{P}_\varepsilon^d(\mathcal{X}, P)} T_Q\right) + \sum_{Q \in \mathcal{P}_\varepsilon^d(\mathcal{X}, P): T_Q \cap A_{\alpha,\beta}^d = \emptyset} P(T_Q) \\
& \leq o(1) + \sum_{Q \in \mathcal{P}_\varepsilon^d(\mathcal{X}, P): T_Q \cap A_{\alpha,\beta}^d = \emptyset} \frac{P(T_{\sigma(Q)})}{1 + O(\varepsilon) + O(d^{-1})} \\
& \leq o(1) + \sum_{Q: T_Q \subset A_{\alpha,\beta}^d} \frac{P(T_Q)}{1 + O(\varepsilon) + O(d^{-1})} \\
& = o(1) + \frac{P(A_{\alpha,\beta}^d)}{1 + O(\varepsilon) + O(d^{-1})}
\end{aligned}$$

Where we have used the AEP (Lemma 2) in the second inequality. The result now follows by rearranging the terms above, taking the limit as  $d \rightarrow \infty$ , and noting that  $\varepsilon > 0$  can be taken to be arbitrarily small. ■

#### ACKNOWLEDGMENTS

We would like to thank Yuriy Reznik for pointing out Khodak's paper. We are also grateful to the anonymous reviewers for their insightful comments and suggestions that have helped improve the presentation of the paper.

#### REFERENCES

- [1] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.
- [2] D.A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. of the I.R.E.*, pp. 1098–1102, September 1952.
- [3] B. P. Tunstall, *Synthesis of Noiseless Compression Codes*, Ph.d. dissertation, Georgia Inst. Tech., Atlanta, GA, 1967.
- [4] G.L. Khodak, "Delay-redundancy relation of VB-encoding (in Russian)," *All-union Conference on Theoretical Cybernetics, Novobirsk*, 1969.
- [5] W. Szpankowski, "Asymptotic average redundancy of Huffman (and other) block codes," *IEEE Trans. on Info. Theory*, vol. 46, no. 7, Nov 2000.
- [6] M. Drmota, Y. Reznik, S.A. Savari, and W. Szpankowski, "Precise asymptotic analysis of the Tunstall code," in *Proc. of the International Symposium on Information Theory*, 2006, pp. 2334–2337.
- [7] G.L. Khodak, "Bounds of redundancy estimates for word-based encoding of sequences produced by a Bernoulli source (Russian)," *Probl. Pered. Inform.*, vol. 8, pp. 21–32, 1972.
- [8] Y. Bugeaud, M. Drmota, and W. Szpankowski, "On the construction of (explicit) Khodak's code and its analysis," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5073–5086, Nov 2008.
- [9] T. J. Tjalkens and F. M. J. Willems, "Variable to fixed-length codes for Markov sources," *IEEE Trans. Info. Theory*, vol. IT-33, no. 2, pp. 246–257, March 1987.
- [10] N. Abramson, *Information Theory and Coding*, McGraw-Hill, New York, 1963.
- [11] F. Jelinek, *Probabilistic Information Theory*, McGraw-Hill, New York, 1968.
- [12] J. Rissanen and G. G. Langdon Jr., "Arithmetic coding," *IBM Journal of research and development*, vol. 23, no. 2, pp. 149–162, 1979.
- [13] R. M. I. A. Witten, Neal, and Cleary J. G., "Arithmetic coding for data compression," *Commun. ACM*, vol. 30, no. 6, pp. 520–540, June 1987.
- [14] R.G. Gallager, *Lecture Notes (unpublished)*, 1991.
- [15] S.A. Savari and R.G. Gallager, "Arithmetic coding for finite-state noiseless channels," *IEEE Trans. Info. Theory*, vol. 40, pp. 100 – 107, 1994.
- [16] O. Shayevitz, R. Zamir, and M. Feder, "Bounded expected delay in arithmetic coding," in *Proc. of the International Symposium on Information Theory*, July 2006.
- [17] O. Shayevitz, E. Meron, M. Feder, and R. Zamir, "Bounds on redundancy in constrained delay arithmetic coding," in *Proc. of the Data Compression Conference*, 2007, pp. 133–142.
- [18] E. Meron, O. Shayevitz, M. Feder, and R. Zamir, "A lower bound on the redundancy of arithmetic-type delay constrained coding," in *Proc. of the Data Compression Conference*, 2008.
- [19] S. Savari and R.G. Gallager, "Arithmetic coding for finite-state noiseless channels," Tech. Rep. LIDS-P ; 2143, MIT, 1992.
- [20] A. Moffat, R. M. Neal, and I. H. Witten, "Arithmetic coding revisited," *ACM Trans. Inf. Syst.*, vol. 16, no. 3, pp. 256–294, July 1998.
- [21] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Trans. Info. Theory*, vol. IT-14, pp. 490 – 501, May 1968.
- [22] I. Csiszár and J. Körner, *Information theory : Coding theorems for discrete memoryless systems*, Cambridge University Press, 2nd Edition, 2011.
- [23] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Sympos. Math. Stat. and Prob.*, 1960, vol. 1, pp. 547–561.
- [24] O. Shayevitz, "On Rényi measures and hypothesis testing," in *Proc. of IEEE International Symposium on Information Theory*, July 2011, pp. 894–898.
- [25] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Info. Theory*, vol. IT-24, pp. 530 – 536, Sept. 1978.
- [26] S. A. Savari, "Renewal theory and source coding," *Proc. of the IEEE*, vol. 88, no. 11, pp. 1692 –1702, Nov 2000.
- [27] M.J. Weinberger, A. Lempel, and J. Ziv, "A sequential algorithm for the universal coding of finite memory sources," *IEEE Trans. Info. Theory*, vol. 38, no. 3, pp. 1002 – 1014, May 1992.