Universal Polarization for Processes with Memory

Boaz Shuval, Student Member, IEEE, Ido Tal, Senior Member, IEEE

Abstract—A transform that is universally polarizing over a set of channels with memory is presented. Memory may be present in both the input to the channel and the channel itself. Both the encoder and the decoder are aware of the input distribution, which is fixed. However, only the decoder is aware of the actual channel being used. The transform can be used to design a universal code for this scenario. The code is to have vanishing error probability when used over any channel in the set, and achieve the infimal information rate over the set. The setting considered is, in fact, more general: we consider a set of processes with memory. Universal polarization is established for the case where each process in the set: (a) has memory in the form of an underlying hidden Markov state sequence that is aperiodic and irreducible, and (b) satisfies a ‘forgetfulness’ property. Forgetfulness, which we believe to be of independent interest, occurs when two hidden Markov states become approximately independent of each other given a sufficiently long sequence of observations between them. We show that aperiodicity and irreducibility of the underlying Markov chain is not sufficient for forgetfulness, and develop a sufficient condition for a hidden Markov process to be forgetful.

Index Terms—Polar codes, universal polarization, universal codes, channels with memory, hidden Markov processes

I. INTRODUCTION

IMPERFECT channel knowledge characterizes many practical communication scenarios. There are various models for imperfect channel knowledge; see [1] for a comprehensive discussion. We consider the scenario where the decoder has full channel information, but the encoder is only aware of a set to which the actual channel belongs. Both the encoder and the decoder are aware of the input distribution, which is fixed. We wish to build a polarization-based code that is universal over the set: it achieves vanishing error probability for any channel in the set, and its rate approaches the infimal information rate over all channels in the set.

In fact, this work tackles a more general setting. The universal construction in this paper applies both to channel coding and source coding scenarios. However, to keep the introduction focused, we concentrate on a channel-coding scenario. We wish to design polarization-based codes that achieve vanishing error probability over a set of channels with memory. The input distribution to all channels in the set is fixed and known at the encoder and decoder. The encoder only knows that the channel belongs to the set, while the decoder is aware of the actual channel used. Examples of channels with memory are finite-state channels, input-constrained channels, and intersymbol-interference channels. We show a polar coding construction that approaches the infimal information rate among the set of channels under successive-cancellation decoding, provided that every input-output process in the set satisfies some mild technical constraints. This construction achieves vanishing error probability over all processes in this set with the same exponent as Arikan’s polar codes [2], [3].

The study of polar coding for a class of memoryless channels with full channel knowledge at the decoder was first considered in [4]. Hassani et al. showed that Arikan’s polar codes [2], under successive-cancellation decoding, cannot achieve the compound capacity [5] of a set of binary-input, memoryless, and symmetric (BMS) channels. In [6, Proposition 7.1] it was shown that polar codes are universal over a set of BMS channels if optimal decoding is employed. Thus, the non-universality exhibited in [4] is an artifact of using successive-cancellation decoding. Nevertheless, coding methods that are based on polarization have been shown to yield universal codes.

In [7], Hassani and Urbanke present two designs based on Arikan’s polar codes that achieve universality over a set of BMS channels. Their first construction combines Arikan’s polar codes and Reed-Solomon codes designed for an erasure channel. Their second construction may be viewed as a two-stage method. In the first stage, one forms several Arikan polar codes, in which identical channels are combined recursively. In the second stage, different channels are combined to obtain universality.

Şaşoğlu and Wang [8] presented another universal polar coding construction for BMS channels. Their construction is also a recursive two-stage method. The first stage, called the slow stage, transforms multiple channel-uses into ones that universally have high-entropy and ones that universally have low-entropy. The second stage, invoked once sufficient polarization is obtained, combines the channels that are universally low-entropy using Arikan’s polar codes to yield vanishing error probability. The construction presented in this paper is a simplified variation of the Şaşoğlu-Wang construction.

We briefly mention other works concerning universality of polar codes. Universal polar codes for families of ordered BMS channels or memoryless sources, with full decoder side information, was considered in [9]. See also [10] for the case of universal polar source codes, with specialization to the binary case. Universal source polarization was studied in [11], in which polar-based codes were used to compress a memoryless source to be losslessly recovered by multiple users, each observing different local side information on the source sequence. Finally, universal polar coding for certain classes of BMS channels with channel knowledge at the encoder was considered in [12].

We present our universal construction in Section III. It consists of two stages, a slow stage, described in Section III-B, followed by a fast stage, described in Section III-C. Both stages are recursive and use Arikan transforms as building blocks. The fast stage consists of multiple applications of Arikan transforms as in the seminal paper [2]. The slow stage uses Arikan transforms in a different manner. Properties of the slow stage, as well as a variation of it that will be useful for our proof of universality, are presented in Section IV. When used over a set of BMS channels and specialized appropriately,
Polar codes were shown to achieve vanishing error probability for processes with memory in [13] and [14]. It was shown in [13] that a large class of processes with memory polarizes under Arıkan’s polar transform. This result extended Şasoğlu’s earlier findings in [6, Chapter 5]. It was further shown in [13] that the Bhattacharyya parameter polarizes fast to 0 for this class. Later, it was shown in [14] that for processes with an underlying hidden Markov structure, the Bhattacharyya parameter also polarizes fast to 1. Combined, the results of [13] and [14] enable information-rate-achieving polar codes for such processes with memory. A practical, low-complexity, decoding algorithm for processes with memory with an underlying hidden Markov structure was described in [15] and [16]. This algorithm is a variation of successive-cancellation decoding that takes into account the hidden state.

One drawback of polar codes for processes with memory using the strategy above is that they must be tailored for the process. For example, to design a polar code for a channel with intersymbol interference, one must know the exact transfer function of the channel. In a practical scenario, it is reasonable to assume that the decoder has full channel knowledge, obtained, for example, by channel estimation based on a reference sequence [17]. However, the assumption that the encoder also has full channel knowledge before transmission may be unrealistic. This is where universal polar codes come into play.

In the universal setting we consider, the encoder has partial information: it knows that the process belongs to some set of processes with memory. The exact process is known only to the decoder, at the time of decoding. The encoder must employ a code that will enable vanishing error probability no matter which process in the set is used. We wish to design a universal code with the highest possible rate over the entire set. Thus, the code is to approach the infimal information rate over the entire set.

This is indeed what we achieve in this work. We show that our polarization-based construction is universal over sets of processes with memory. We prove universality when the sets contain processes with memory that satisfy two technical constraints, presented in detail in Section V-A. Briefly, the processes have an underlying hidden finite-state Markov structure that is regular (aperiodic and irreducible); and they have a property we call forgetfulness, which we believe is of independent interest.

Forgetfulness is a property we now describe informally. In a hidden Markov process, we are given a sequence of observations that are known to be probabilistic functions of some Markov chain called the state process. The process is called forgetful if, given a long-enough sequence of observations, the state at the time of the first observation and the state at the time of the last observation become approximately independent. Surprisingly, regularity of the underlying Markov chain is not sufficient to ensure forgetfulness. We note that forgetfulness was not required in the non-universal setting of [13], [14], yet in our proof of the universal case it plays a key role.

Hochwald and Jelenković [18] considered a property similar to forgetfulness under the restrictive assumption that there is a positive probability of transitioning between any two states in one step. Leveraging ideas from Kaijser [19], we lift this restrictive assumption and prove, in Sections VII and VIII, a sufficient condition for forgetfulness of a hidden Markov model. This condition, which we call Condition K, takes into account both the transition matrix of the state process as well as the probabilistic function that generates the observations. We show that Condition K yields exponentially fast forgetfulness. Specifically, we use mutual information as a measure for independence, and show that under Condition K, the mutual information between the states at the beginning and end of a block, given the observations in between, vanishes exponentially fast with the length of the block.

The slow stage of the construction is the one responsible for its universality. The proof of universality is given in Sections V-B and V-C. Low complexity decoding of the universal polar codes is based on the successive-cancellation trellis decoding of [16]; details are given in Section VI.

**Paper Roadmap:** There are several ways to read this paper, with increasing levels of detail. A map of the various paths is shown in Figure 1. All readers are advised to familiarize themselves with the notations and definitions of Section II. In it, we introduce the notion of a symbol/observation pair, which generalizes the concept of a channel and allows for simultaneous description of channel and source coding. Section III is also recommended for all readers, for it introduces the details of the universal construction. At this point, there are several options.

- A practitioner who wishes to understand and implement the construction, without getting bogged down with the proofs, is advised to skip to Section V-A, and read it up to Example 4. This introduces the assumptions.
on the processes for which we can prove universality. Examples 3 and 4 are important as they illustrate that forgetfulness does not follow from regularity (aperiodicity and irreducibility) of the underlying Markov chain. Then, the practitioner may skip straight to the decoding process in Section VI.

- A reader who is interested in understanding why the construction is universal is advised to turn to Sections IV and V after Section III. These sections contain a detailed proof of universality of the construction, provided that one takes on faith that forgetful processes exist.
- A sufficient condition for the existence of forgetful processes is developed in Sections VII and VIII. The interested reader is advised to read them following Section V-A. Sections VII and VIII are written for a general hidden Markov model and may be read independently.

II. NOTATION AND BASIC DEFINITIONS

A discrete set of elements is denoted as a list in braces, e.g., \( \{1, 2, \ldots, L\} \), usually denoted with a calligraphic letter, e.g., \( \mathcal{A} \). The number of elements in a discrete set \( \mathcal{A} \) is denoted by \(|\mathcal{A}|\). We denote \( y_j^k = [y_j \ y_{j+1} \ \cdots \ y_k] \) for \( j < k \). If \( j = k \) then \( y_j^k = y_j \) and if \( j > k \) then \( y_j^k \) is a null vector.

We use boldface to denote vectors, and, unless stated otherwise, vectors are assumed to be column vectors. The transpose of a column vector \( \mathbf{x} \) is the row vector \( \mathbf{x}^T \). The \( i \)th element of a vector \( \mathbf{x} \) is denoted by \( (\mathbf{x})_i \) (usually, and unless stated otherwise, \( (\mathbf{x})_i = x_i \)). Special vectors are the all-ones vector \( \mathbf{1} \), all-zeros vector \( \mathbf{0} \), and the unit vector \( \mathbf{e}_i \), which has 1 in its \( i \)th entry and zero in all other entries. We further define the norm \( \|\mathbf{x}\|_1 = \sum_i |x_i| \).

An inequality involving vectors is assumed to be element-wise. Therefore, if \( a \) is a scalar and \( \mathbf{b} \) is a vector, \( \mathbf{x} \geq a \) implies that \( x_i \geq a \) for all \( i \), and \( \mathbf{x} \geq \mathbf{b} \) implies that \( x_i \geq b_i \) for all \( i \). For two vectors (possibly of different lengths) \( \mathbf{a} \) and \( \mathbf{b} \) we write \( \mathbf{a} \preceq \mathbf{b} \) if there is a one-to-one mapping \( f \) between \( \mathbf{a} \) and \( \mathbf{b} \); usually, \( f \) is clear from the context, so we omit it and simply write \( \mathbf{a} \preceq \mathbf{b} \). The support \( \sigma(\mathbf{x}) \) of a vector \( \mathbf{x} \) is the set of indices \( i \) such that \( x_i \neq 0 \). A vector is said to be nonzero if it has a non-empty support.

Matrices are denoted using capital letters in sans-serif font, e.g., \( \mathbf{M} \). The \( i, j \) element of a matrix \( \mathbf{M} \) is denoted by \( (\mathbf{M})_{i,j} \). The \( i \)th row of \( \mathbf{M} \) is denoted by \( (\mathbf{M})_i \), and the \( j \)th column of \( \mathbf{M} \) is denoted by \( (\mathbf{M})_j \). The identity matrix is denoted by \( \mathbf{I} \). For matrix \( \mathbf{M} \), we denote its set of nonzero rows \( \mathcal{N}_i(\mathbf{M}) \) and its set of nonzero columns \( \mathcal{N}_i(\mathbf{M}) \). The support \( \sigma(\mathbf{M}) \) of a matrix \( \mathbf{M} \) is the set of index pairs \( (i,j) \) such that \( i \in \mathcal{N}_i(\mathbf{M}) \) and \( j \in \mathcal{N}_i(\mathbf{M}) \).

The probability of an event \( A \) is denoted by \( \mathbb{P}(A) \). Random variables are usually denoted using upper-case letters, e.g., \( X \), and their realizations using lower-case letters, e.g., \( x \). The distribution of random variable \( X \) is denoted by \( P_X \). The expectation of \( X \) is denoted by \( \mathbb{E}[X] \). When \( X_n \) is a sequence

1\ A row or column is nonzero if it has at least one nonzero element.

of random variables and \( \mathbf{b} = [b_1 \ b_2 \ \cdots \ b_m] \) is a vector of indices, then \( X_\mathbf{b} = (X_{b_1}, X_{b_2}, \ldots, X_{b_m}) \).

Let \( X \) and \( Y \) be two discrete random variables taking values in alphabets \( X \) and \( Y \), respectively. We define \( H(X) \), the entropy of \( X \), and \( H(X|Y) \), the conditional entropy of \( X \) given \( Y \), by

\[
H(X) = -\sum_{x \in X} P_X(x) \log P_X(x),
\]

\[
H(X|Y) = -\sum_{y \in Y, x \in X} P_{X,Y}(x,y) \log P_{X|Y}(x|y),
\]

where we follow the usual convention that \( 0 \cdot \log 0 = 0 \). Logarithms are base 2 unless stated otherwise. The binary entropy function \( h_2 : [0,1] \to [0,1] \) is defined by

\[
h_2(x) = -x \log x - (1-x) \log(1-x).
\]

The mutual information between \( X \) and \( Y \), denoted \( I(X;Y) \) is defined by

\[
I(X;Y) = H(X) - H(X|Y).
\]

Let \( Q \) be an additional discrete random variable; the conditional mutual information of \( X \) and \( Y \) given \( Q \) is \( I(X;Y|Q) = H(X|Q) - H(X|Y,Q) \).

The following variation of the data processing inequality will be useful. Let \( X, Y, Q, W \) be four random variables. We introduce the notation \( X \prec (Y, Q) \prec W \) whenever \( X \) and \( W \) are independent given \( Y \) and \( Q \). We then have the following variation of the data processing inequality:

\[
X \prec (Y, Q) \prec W \Rightarrow I(X;Y|Q) \geq I(X;W|Q).
\]

Indeed, on the one hand, \( I(X;(Y, W)|Q) = I(X;Y|Q) + I(X;W|Y,Q) = I(X;Y|Q) \), where the last equality is by conditional independence. On the other hand \( I(X;(Y, W)|Q) = I(X;W|Q) + I(X;Y|W, Q) \geq I(X;W|Q) \), since mutual information is nonnegative.

The following definition generalizes the concept of a channel. This generalization allows us to describe polarization transforms for channel coding and source coding in one fell swoop.

**Definition 1** (s/o-pair). A symbol-observation pair, or s/o-pair in short, is a pair of dependent random variables \( X \) and \( Y \). The random variable \( X \) is called the symbol and the random variable \( Y \) is called the observation. We use the notation \( X \rightarrow Y \) to denote an s/o-pair whose symbol is \( X \) and whose observation is \( Y \). The joint distribution of the s/o-pair is given by \( P_{X,Y}(x,y) = P_X(x)P_{Y|X}(y|x) \). The conditional entropy of an s/o-pair \( X \rightarrow Y \) is \( H(Y | X) \).

We emphasize that an s/o-pair is specified using the joint distribution of \( X \) and \( Y \). This is in contrast to a channel that is specified using only the conditional distribution of the output given its input. A channel with input \( X \) and output \( Y \) becomes an s/o-pair once the input distribution is specified. Another example of an s/o-pair is a source \( X \) with distribution \( P_X(x) \) to be estimated based on observation \( Y \) distributed according to \( P_{Y|X}(y|x) \).

**Definition 2** (s/o-process). A sequence of s/o-pairs \( X_t \rightarrow Y_t \), \( t = 1, 2, \ldots \) is called a symbol-observation process, or s/o-process in short. We use the notation \( X_t \rightarrow Y_t \)
Definition 3 (s/o-block). A sequence of \( N \) consecutive s/o-pairs of an s/o-process is called an s/o-block. We use the notation \( X_i^N \to Y_i^N \). An s/o-block has a natural indexing: \( X_j \to Y_j \) is s/o-pair \( j \) of s/o-block \( X_i^N \to Y_i^N \). The joint distribution of an s/o-block is given by \( P_{X_i^N,Y_i^N}(x_i^N,y_i^N) = P_{X_i^N}(x_i^N)P_{Y_i^N|X_i^N}(y_i^N|x_i^N) \).

Generally, the s/o-pairs in an s/o-block are dependent; that is, there is memory in the process. In this paper, we assume that s/o-processes are stationary. In particular, this implies that for an s/o-block \( X_i^N \to Y_i^N \), the s/o-pairs \( X_j \to Y_j \) are identically distributed for all \( i \).

The conditional entropy rate of a stationary s/o-process \( X_\ast \to Y_\ast \) is

\[
\mathcal{H}(X_\ast|Y_\ast) = \lim_{N \to \infty} \frac{1}{N} \mathcal{H}(X_i^N|Y_i^N) = \lim_{N \to \infty} \frac{1}{N} \mathcal{H}(X_i^N) - \lim_{N \to \infty} \frac{1}{N} \mathcal{H}(Y_i^N).
\]

The limits on the right-hand side exist due to stationarity (see, e.g., [20, Theorem 4.2.1]).

For simplicity, we assume throughout that s/o-pairs have binary symbols and that their observations are over a finite alphabet. Extension to the case where symbols are non-binary and that an alphabet of prime size is possible using the techniques of [6, Chapter 3]. This entails replacing modulo-2 addition with modulo-\(|X|\) addition, where \(|X|\) is the symbol alphabet size, and replacing binary entropies with non-binary entropies.

### III. Universal Polar Transform

In this section we describe the universal polar transform, which is based on [8]. The transform described in [8] was used to construct a universal code over memoryless symmetric channels subject to a capacity constraint. In this work, we extend the transform of [8] for s/o-processes with memory.

This section is focused on describing the transform. Properties of the transform and proof of its universality are presented in Sections IV and V. The decoding operation is described in Section VI.

**A. Overview of the Transform**

In this section, we provide a general overview of the universal polar transform. It is a type of H-transform, a concept that we now define.

**Definition 4 (H-transform).** A one-to-one and onto mapping \( f \) between two symbol vectors of length \( N \) is called an H-transform.

Moreover, when we say that s/o-block \( X_i^N \to Y_i^N \) is transformed to s/o-block \( F_i^N \to G_i^N \) by H-transform \( f \), we mean that:

1. \( F_i^N = f(X_i^N) \);
2. \( G_i = (F_i^{-1}; Y_i^N) \), for any \( i \).

**Example 1.** Arikan’s polar codes [2] are based on H-transforms. In this case, the mapping \( f \) is given by \( F_i^N = f(X_i^N) = B_N G_2^\otimes x_i^N \), where \( N = 2^N \), \( B_N \) is the \( N \times N \) bit-reversal matrix, \( G_2 = \begin{bmatrix} 1 & 1 \end{bmatrix} \), and \( \otimes \) denotes a Kronecker product.

The name “H-transform” is motivated by the equality

\[
H(X_i^N|Y_i^N) = H(F_i^N|Y_i^N) = \sum_{i=1}^{N} H(F_i|G_i).
\]

The right-most equality follows from the chain rule for entropies and the definition of \( G_i \). Typically, the \( f \) of an H-transform is defined recursively.

Consider an s/o-block \( X_i^N \to Y_i^N \), with H-transform \( F_i^N \to G_i^N \). We wish to recover the symbols \( X_i^N \) from the observations \( Y_i^N \). We denote the recovered symbols with a hat, \( \hat{\cdot} \). That is, \( \hat{X}_i^N = \Phi(Y_i^N) \), where \( \Phi(\cdot) \) is the algorithm for recovery. We assess \( \Phi \) by its error probability, \( P(\hat{X}_i^N \neq X_i^N) \). H-transforms, thanks to (3), naturally give rise to a sequential algorithm called successive cancellation.

Rather than computing \( \hat{X}_i^N \) from \( Y_i^N \) directly, we may compute \( \hat{F}_i^N \) from \( Y_i^N \). By the properties of the H-transform, there exists a mapping \( f \), with inverse \( f^{-1} \), such that \( \hat{X}_i^N = f^{-1}(\hat{F}_i^N) \). Any algorithm for recovering \( \hat{F}_i^N \) from \( Y_i^N \) is equivalent to an algorithm for recovering \( \hat{X}_i^N \) from \( Y_i^N \). For, if \( \hat{F}_i^N = \Phi(Y_i^N) \) we can define \( \hat{X}_i^N = f^{-1}(\hat{F}_i^N) = f^{-1}(\Phi(Y_i^N)) \) and vice versa. Since \( P(\hat{F}_i^N \neq \hat{F}_i^N) = P(\hat{X}_i^N \neq X_i^N) \), we concentrate on an algorithm to recover \( \hat{F}_i^N \).

One approach is to compute \( \hat{F}_i^N \) sequentially as follows. Let \( \Phi_i \) be a maximum-likelihood decoder of \( F_i \) from \( G_i \). Compute \( \hat{F}_i = \Phi_i(\hat{G}_i) \), where \( \hat{G}_i = G_i = Y_i^N \); then, assuming that \( \hat{F}_i = F_i \), form \( \hat{G}_2 = (\hat{F}_i; Y_i^N) \) and compute \( \hat{F}_2 = \Phi_2(\hat{G}_2) \), and so on, culminating with \( \hat{F}_N = \Phi_N(\hat{G}_N) \). This is tantamount to the successive-cancellation decoding described in [2], and we will use the name “successive cancellation” to describe this algorithm.

It is well known [6, Proposition 2.1] that the error probability of recovering \( \hat{F}_i^N \) sequentially from \( \hat{G}_i^N \) using successive cancellation as described above is the same as if a genie had replaced \( G_i \) with \( G_i \) at every step. That is,

\[
P \left( \left( \left( \Phi_i(\hat{G}_i) \right)_{i=1}^{N} \neq (F_i)_{i=1}^{N} \right) \right) = P \left( \left( \left( \Phi_i(\hat{G}_i) \right)_{i=1}^{N} \neq (F_i)_{i=1}^{N} \right) \right).
\]

To see this, observe that if \( \Phi_i(\hat{G}_i) = F_i \) for all \( i < i_0 \) and \( \Phi_{i_0}(\hat{G}_{i_0}) \neq F_{i_0} \) then we must also have \( \Phi_i(\hat{G}_i) = F_i \) for all \( i < i_0 \) and \( \Phi_{i_0}(\hat{G}_{i_0}) \neq F_{i_0} \). Therefore, when assessing the performance of successive cancellation, we may assume that at step \( i \), \( G_i \) (in contrast to \( \hat{G}_i \)) is known.

**Definition 5 (Monopolarizing H-transform).** Let \( \eta > 0 \) and let \( \mathcal{L}, \mathcal{M} \subseteq \{1, 2, \ldots, N\} \) be two index sets. An H-transform \( f \) is \((\eta, \mathcal{L}, \mathcal{M})\)-monopolarizing for a family of s/o-processes if for any s/o-block \( X_i^N \to Y_i^N \) in the family, either \( H(F_i|G_i) \leq \eta \) for all \( i \in \mathcal{L} \) or \( H(F_i|G_i) \geq 1 - \eta \) for all \( i \in \mathcal{M} \), where s/o-block \( F_i^N \to G_i^N \) denotes the transformed s/o-block.

Monopolarizing H-transforms are useful because they make the process of recovering \( \hat{F}_i \) from \( G_i \) very easy whenever \( H(F_i|G_i) \approx 0 \), because then \( F_i \) is approximately a deterministic function of \( G_i \). On the other hand, if \( H(F_i|G_i) \approx 1 \) we know that \( F_i \) is essentially a result of a uniform coin flip, independent of \( G_i \).

The universal transform is a moniker for a family of H-transforms with increasing lengths. It comprises two stages: a slow polarization stage and a fast polarization stage. Each
is an H-transform that is constructed recursively. Our goal is to show that, as the blocklength increases, they become monopolarizing.

Recursive construction of an H-transform begins with an initial H-transform $f_0$ of length $N_0$. Then, at step $n+1$ we take step-$n$ H-transforms of consecutive symbol vectors to generate a step-$(n+1)$ H-transform of a single, larger, symbol vector. A typical case is as follows. Let $f_n$ be an H-transform of length $N_n$, that results from step $n$, and let $\varphi_{n+1}$ be a one-to-one and onto mapping from two length-$N_n$ vectors to a vector of length $N_{n+1} = 2N_n$. Apply $f_n$ to two consecutive symbol vectors: $U^N_1 = f_n(X^N_1)$ and $V^N_1 = f_n(X^N_{n+1})$. Then, form $F^N_{n+1} = \varphi_{n+1}((U^N_1, V^N_1)) = f_{n+1}(X^N_{n+1})$.

A basic building block is the Arıkan transform [2], illustrated in Figure 2. It operates on two input symbols: input-I: $U$ (with observation $Q$) and input-II: $V$ (with observation $R$) and transforms them to two new symbols: a ‘$-$’ symbol $F_1$ (with observation $G_1$) and a ‘$+$’ symbol $F_2$ (with observation $G_2$), where $F_1 = U + V$, $G_1 = (Q, R)$ and $F_2 = V$, $G_2 = (F_1, Q, R)$. Schematically, the Arıkan transform is as follows:

$$\begin{array}{l}
I: U \rightarrow Q \\
\quad \quad \quad \chi \quad \rightarrow (Q, R) \\
\quad \quad \quad F_1 \quad \rightarrow G_1 \\
\quad \quad \quad F_2 \quad \rightarrow G_2 \\
\end{array}$$

It is evident that an Arıkan transform is an H-transform of length 2.

For an Arıkan transform, we obtain

$$H(F_1|G_1) + H(F_2|G_2) = H(F_1^Q|Q, R) = H(U, V|Q, R) \leq H(U|Q) + H(V|R).$$

The inequality is because the s/o-pairs $U \mapsto Q$ and $V \mapsto R$ are generally dependent. Informally, Arıkan transforms facilitate polarization if one can show that $H(F_1|G_1) \geq \max\{H(U|Q), H(V|R)\}$ and that the inequality is strict unless either $H(U|Q)$ or $H(V|R)$ is extremal. This was the strategy of obtaining polarization for standard (Arıkan’s) polar codes, with and without memory. See, for example, [2], [6], [13]. We will also pursue such a strategy.

### B. Slow Polarization Stage

In this subsection we describe the slow polarization stage. We will focus on describing a slow stage transform called a basic slow transform (BST). It is an extension of the transform shown in [8, Section II].

The basic slow transform is constructed recursively. We call each step in the construction a level. Each level is an H-transform of length $N_n = 2L_n + M_n$. We will specify how to compute $L_n$ and $M_n$ later in (8). We call the transformed s/o-block a level-$n$ block.

We define the following index sets for a level-$n$ block, $n \geq 0$. See Figure 3 for an illustration.

- $[\text{lat}_1(n)] \triangleq \{ i \mid 1 \leq i \leq L_n \}$
- $[\text{lat}_2(n)] \triangleq \{ i \mid L_n + M_n + 1 \leq i \leq N_n \}$
- $[\text{lat}(n)] \triangleq [\text{lat}_1(n)] \cup [\text{lat}_2(n)]$
- $[\text{med}_1(n)] \triangleq \{ i \mid i = L_n + 2k - 1, 1 \leq k \leq M_n/2 \}$
- $[\text{med}_2(n)] \triangleq \{ i \mid i = L_n + 2k, 1 \leq k \leq M_n/2 \}$
- $[\text{med}(n)] \triangleq [\text{med}_1(n)] \cup [\text{med}_2(n)]$

In words, the sets $[\text{lat}_1(n)]$ and $[\text{lat}_2(n)]$ are, respectively, the first $L_n$ and last $L_n$ indices in a level-$n$ block. Then, the remaining $M_n$ indices alternate between $[\text{med}_1(n)]$ and $[\text{med}_2(n)]$, starting with $[\text{med}_1(n)]$ and ending with $[\text{med}_2(n)]$.

We classify symbols in an s/o-block according to their indices as follows:

- $i \in [\text{lat}(n)] \Rightarrow \text{symbol } i \text{ is lateral}$
- $i \in [\text{med}(n)] \Rightarrow \text{symbol } i \text{ is medial}$

We will sometimes classify s/o-pairs based on the classification of the indices. For example, we say that s/o-pair $i$ is lateral if symbol $i$ is lateral.

The construction is initialized with integer parameters $L_0$ and $M_0$. We assume that $M_0$ is even.

- The parameter $L_0$ determines, informally, “how much memory” in the s/o-process the transform can handle; see Section V for more details. For a memoryless process, it may be set to 0.
- The parameter $M_0$ has a dual role:
  - Informally, it is set large enough so that two s/o-pairs that are $M_0$ time-indices apart may be considered almost independent. See Section V for more details.
  - It controls the fraction of medial symbols in an s/o-block. See Lemma 2 for details.

The initial step $f_0$, which generates a level-0 block, is an H-transform of length $N_0 = 2L_0 + M_0$. We set $f_0$ as the identity mapping. Thus, the initial step transforms an s/o-block $X^N_1 \rightarrow Y^N_1$ into an s/o-block $X^N_1 \rightarrow Y^N_1$. We will now construct a level-$(n+1)$ BST from two level-$n$ BSTs. Denote by $f_n$ a BST of length $N_n$. We will define $f_{n+1}$ using a one-to-one and onto mapping $\varphi_{n+1}$ from two length-$N_n$ vectors to a single length-$N_{n+1} = 2N_n$ vector. The mapping $\varphi_{n+1}$ is defined in (9) and (10) below.

The BSTs of the two consecutive level-$n$ s/o-blocks are

$$U^N_1 = f_n(X^N_1), \quad Q_i = (U^1_{i-1}, Y^N_n), \quad 1 \leq i \leq N_n.$$  

$$V^N_1 = f_n(X^{2N}_n), \quad R_i = (V^1_{i-1}, Y^{2N}_n), \quad 1 \leq i \leq N_n.$$  

This is not necessary, and it is possible to initialize the construction with odd $M_0$. However, assuming that $M_0$ is even ensures that the index sets defined in (4) hold also for $n = 0$. 

---

[Note: The image represents an Arıkan transform diagram with inputs and outputs labeled.]
Denoting $N_{n+1} = 2N_n$, we obtain the level-$(n+1)$ transformed s/o-block

$$F_{1}^{N_{n+1}} = \varphi_{n+1}(U_{1}^{N_n}, V_{1}^{N_n}) = f_{n+1}(X_{1}^{N_{n+1}}),$$

$$G_i = (F_{i-1}^{N_{n+1}}, Y_{i}^{N_{n+1}}), \quad 1 \leq i \leq N_{n+1}. \quad (7b)$$

The level-$(n+1)$ block is of length $N_{n+1} = 2L_{n+1} + M_{n+1}$, where

$$L_{n+1} = 2L_n + 1 \quad (8a)$$

$$M_{n+1} = 2(M_n - 1). \quad (8b)$$

Indeed, $N_{n+1} = 2L_{n+1} + M_{n+1} = 2(2L_n + M_n) = 2N_n$.

**Remark 1.** Observe that $L_n$ is odd and $M_n$ is even for any $n \geq 1$. Therefore, for any $n \geq 1$, the set $[\text{med}(n)]$ is the set of even indices of $[\text{med}(n)]$ and the set $[\text{med}(n)]$ is the set of odd indices of $[\text{med}(n)]$.

Lateral symbols of a level-$(n+1)$ block are formed by renaming symbols of level-$n$ s/o-pairs, as follows:

$$i \in [\text{lat}(n+1)] \Rightarrow F_i = \begin{cases} U_{j}, & i = 2j, \\ V_{j}, & i = 2j + 1. \end{cases} \quad (9)$$

This is illustrated in Figure 4. Observe that all lateral symbols of the level-$n$ blocks become lateral symbols of the level-$(n+1)$ block. Additionally, note that, by (4), (8), and (9), two medial symbols of the level-$n$ blocks become lateral symbols of the level-$(n+1)$ block:

$$F_{L_{n+1}} = F_{2(L_{n+1})-1} = U_{L_{n+1}}$$

and

$$F_{L_{n+1}+M_{n+1}} = F_{2(L_{n+1}+M_n)} = V_{L_{n+1}+M_n}.$$
Definition 6 (Ancestors and Base-ancestors). An Arıkan transform — see Figure 2 — maps two symbols, \( U \) and \( V \), into two transformed symbols, \( F_1 \) and \( F_2 \). Medial symbols are generated by Arıkan transforms, as evident by Figure 5 and (10). Let \( i = 2j \in [\text{med}(n+1)] \). Then, \( i+1 \in [\text{med}(n+1)] \) as well, see (4) and Remark 1. Medial symbols \( F_1 \) and \( F_{i+1} \), by (10), are generated by an Arıkan transform of \( U_{j+1} \) and \( V_j \). Symbol \( U_{j+1} \) is in the first level-\( n \)-block and symbol \( V_j \) is in the second level-\( n \)-block. Hence, we define the (immediate) ancestors of both medial symbols \( F_i \) and \( F_{i+1} \) as \( U_{j+1} \) and \( V_j \). Since the immediate ancestors are of level \( n \), we may also call them level-\( n \) ancestors.

Each medial symbol of level \( n \), in turn, has two level-(\( n - 1 \)) medial symbols as its immediate ancestors, see the discussion following (11), Thus, we say that a medial symbol in level \( n + 1 \) has four level-(\( n - 1 \)) ancestors, all medial symbols from four different level-(\( n - 1 \)) blocks. Continuing in this manner, a level-(\( n + 1 \)) symbol has \( 2^{n+1} \) level-0 ancestors, all medial symbols from \( 2^{n+1} \) different level-0 blocks. The level-0 ancestors of an symbol are called base-ancestors.

Equations (9) and (10) form a one-to-one and onto mapping from \( (U_1^{N_n}, V_1^{N_n}) \) to \( F_1^{2N_n+1} \). We define the function \( \varphi_{n+1} \) of (7) using these equations. While the level-(\( n + 1 \)) BST is completely specified by (7), the following lemma provides a direct method of computing \( G_1^{N_{n+1}} \) from \( Q_1^{N_n} \) and \( R_1^{N_n} \).

**Lemma 1.** Consider the BST defined by (7), where \( \varphi_{n+1} \) is defined according to (9) and (10). Then, for any \( n \geq 0, \)

\[
i \in [\text{lat}(n+1)] \Rightarrow G_i \equiv \begin{cases} (Q_j, R_j), & i = 2j - 1, \\
(Q_{j+1}, R_j), & i = 2j \neq 2N_n, \\
(F_{i-1}, Q_{N_n}, R_{N_n}), & i = 2N_n \end{cases}
\]

and

\[
i \in [\text{med}(n+1)] \Rightarrow G_i \equiv \begin{cases} (Q_{j+1}, R_j), & i = 2j, \\
(F_{i-1}, Q_{j+1}, R_j), & i = 2j + 1. \end{cases}
\]

**Proof:** By construction, for \( 1 \leq j \leq N_n \), we have

\[
Q_j = (U_{1}^{j-1}, Y_{1}^{N_n}), ~ R_j = (V_{1}^{j-1}, Y_{N_n}^{2N_n}).
\]

Since

\[
G_i = (F_{1}^{i-1}, Y_{2N_n}^{N_n}),
\]

we need only show that there is a one-to-one mapping between the non-\( F \) portions of the right-hand-sides of (12) and (13) to \( F_{1}^{2i-1} \). We proceed in cases, based on the index \( i \) in the level-(\( n+1 \)) block.

*Case 1*: \( i \in [\text{lat}_1(n+1)] \) — the first half of the lateral set, see (4a).

In this case, to show (12) it suffices to establish

\[
F_1^{i-1} \equiv \begin{cases} (U_{1}^{j-1}, V_{1}^{j-1}), & i = 2j - 1, \\
(U_{1}^{j-1}, V_{1}^{j-1}), & i = 2j. \end{cases}
\]

By (9), if \( i = 2j - 1 \) we have \( F_1^{i-1} \equiv (U_{1}^{j-1}, V_{1}^{j-1}) \). If \( i = 2j \) then \( F_1^{i-1} \equiv (U_{1}^{j}, V_{1}^{j-1}) \). Thus, (14) holds for any \( i \in [\text{lat}_1(n+1)] \).

*Case 2*: \( i \in [\text{med}(n+1)] \) — the medial set, see (4f).

In this case, to show (13) it suffices to establish

\[
F_1^{i-1} \equiv \begin{cases} (U_{1}^{j}, V_{1}^{j-1}), & i = 2j, \\
(F_{i-1}, U_{1}^{j-1}, V_{1}^{j-1}), & i = 2j + 1. \end{cases}
\]

By (8a), if \( i \) is the first medial index, \( i = L_{n+1} + 1 = 2(L_n + 1) \). Hence, \( i - 1 \) is odd and lateral, so by (9), \( F_1^{i-1} \equiv (U_{1}^{j-1}, V_{1}^{L_n}) \), and trivially \( F_1^{i} \equiv (F_i, U_{1}^{j-1}, V_{1}^{L_n}) \). This implies (15) for the first two medial indices. We continue by induction. Assume that for \( i = 2j \in [\text{med}(n+1)] \) we have \( F_1^{2j-1} \equiv (U_{1}^{j-1}, V_{1}^{j-1}) \). Trivially, \( F_1^{2j} \equiv (F_{2j}, U_{1}^{j-1}, V_{1}^{j-1}) \); hence (15) holds for \( i + 1 \) as well. By (10),

\[
F_1^{2(i+1)-1} \equiv (F_1^{2i-1}, F_{2j}, F_{2j+1})
\]

\[
\equiv (F_1^{2i-1}, U_{1}^{j-1}, V_{1}^{j}),
\]

\[
\equiv (U_{1}^{j}, V_{1}^{j-1}),
\]

where for the last equivalence we used the induction assumption. This implies (15) for \( i + 2 \).

Observe that when \( i = 2(L_n + M_n - 1) \) \( \equiv [\text{med}(n+1)] \), that is, when \( i \) is the last even index in \( [\text{med}(n+1)] \), then \( i + 2 \) is the first lateral index in \( [\text{lat}_2(n+1)] \). Equation (16) still holds for \( i + 2 = 2(L_n + M_n) \).

*Case 3*: \( i \in [\text{lat}_2(n+1)] \) — the second half of the lateral set, see (4b).

In this case, to show (12) it suffices to establish

\[
F_1^{i-1} \equiv \begin{cases} (U_{1}^{j-1}, V_{1}^{j-1}), & i = 2j - 1, \\
(U_{1}^{j-1}, V_{1}^{j-1}), & i = 2j \neq 2N_n, \\
(F_{i-1}, U_{1}^{N_n-1}, V_{1}^{N_n-1}), & i = 2N_n. \end{cases}
\]
If \( i \) is the first lateral index in \([\text{lat}_2(n+1)]\), by (8) we have
\[ i = L_{n+1} + M_{n+1} + 1 = 2(L_n + M_n). \]
Thus, by the observation at the end of case 2, \( E(2L_n + M_n) \equiv (U_{2L_n + M_n}, V_{L_n + M_n}) \).
For any other index \( i \in [\text{lat}_2(n+1)] \), by (9) indeed (17) holds, similar to case 1.

We conclude this section by computing the fraction of medial symbols out of all symbols in a level-\( n \) block. To this end, denote
\[ \alpha_n \triangleq \frac{M_n}{2L_n + M_n}. \]

**Lemma 2.** Consider a BST initialized with parameters \( L_0 \geq 0 \) and \( M_0 \), and let \( 0 < \alpha < 1 \).

If
\[ M_0 \geq \frac{2(1 + \alpha L_0)}{1 - \alpha}, \]
then \( \alpha_n \geq \alpha \) for any \( n \geq 0 \).

**Proof:** Plugging \( n = 0 \) in (18) yields \( \alpha_0 = M_0/(2L_0 + M_0) \).
It is straightforward to show from (8) that for any \( n \geq 0 \),
\[ L_n = 2^n(L_0 + (1 - 2^{-n})) \]
\[ M_n = 2^n(M_0 - 2(1 - 2^{-n})). \]

Therefore, recalling that \( N_0 = 2L_0 + M_0 \),
\[ \alpha_n = \frac{M_n}{2L_n + M_n} = \frac{M_0 - 2(1 - 2^{-n})}{2L_0 + M_0} = \frac{\alpha_0 - 2(1 - 2^{-n})}{N_0}. \]

This implies
\[ \alpha_n \geq \alpha_0 - \frac{2}{N_0} = \frac{M_0 - 2}{M_0 + 2L_0}. \]

The right-hand side is an increasing function of \( M_0 \), since its derivative with respect to \( M_0 \) is \( (M_0 - 2)/(M_0 + 2L_0) > 0 \). It remains to find \( m_0 \) such that \((m_0 - 2)/(m_0 + 2L_0) = \alpha \). Then, for any \( M_0 \geq [m_0] \), we will have \( \alpha_n \geq \alpha \). The proof is complete by noting that \( m_0 = 2(1 + \alpha L_0)/(1 - \alpha) \).

**Discussion.** The transform presented in [8], henceforth referred to as the Şasoglu-Wang transform (SWT), is the basis for the BST. The first two levels of the SWT (levels 1 and 2 in [8]) differ from the first two levels of the BST (levels 0 and 1 here). After that, the construction of the two transforms coincide (compare our Figure 5 with [8, Figure 5]). The BST is simpler and more streamlined than the SWT, since all levels of the BST share the same construction. In the memoryless case one can verify that the SWT and BST (with \( L_0 = 0 \)) have the same performance.

We will see in Section V that the BST is effective also for processes with memory, by taking \( L_0 > 0 \).

In Section V we will show that for an appropriate \( \eta \) and family of s/o-processes, the BST is \((\eta, \mathcal{L}, \mathcal{H})\)-monopolizing, with \( \mathcal{L} = [\text{med}_1(n)] \) and \( \mathcal{H} = [\text{med}_2(n)] \), where \( n \) is the level number of the BST. In particular, this implies that \(|\mathcal{L}| = |\mathcal{H}|\), which limits to 1/2 the achievable rates the universal code can yield. It is possible to generate slow stage transforms for which \( \mathcal{L} \) and \( \mathcal{H} \) are of different sizes. One way to achieve this is by chaining multiple BSTs. Details can be found in [8, Section III]; a brief description on how this is accomplished follows.

After a BST, all symbols in \([\text{med}_1(n)]\) have approximately the same conditional entropy; the same is true for all symbols in \([\text{med}_2(n)]\). If \( n \) is sufficiently large, one set will have polarized (e.g., the conditional entropies of s/o-pairs in \([\text{med}_2(n)]\) are all very close to 1). By applying a BST to multiple copies of the other set, we divide its s/o-pairs into two new sets of equal size, one of which will have polarized. This operation can be repeated to tailor the size of the polarized set.

An alternative strategy to modify the sizes of \( \mathcal{L} \) and \( \mathcal{H} \) is to form medial symbols with kernels other than the Arıkan transform. A family of kernels are introduced in [8, Section III]. They can also be adapted to our construction, and we leave this to the interested reader.

### C. Fast Polarization Stage

We will show in Section V that the BST is \((\eta, \mathcal{L}, \mathcal{H})\)-monopolizing for a suitable family of s/o-processes with memory. Moreover, the sets \( \mathcal{L} \) and \( \mathcal{H} \) are predetermined; see the discussion at the end of the previous section. However, even in the memoryless case [8], the speed of polarization is too slow to enable a successive-cancellation decoder to succeed.

Therefore, as in [8], we append a fast polarization stage to the BST that facilitates error-free successive-cancellation decoding.

The fast polarization stage is based on Arıkan’s polar transform [2], which is known to polarize fast also under memory [13, 14]. One strategy to incorporate a fast polarization stage, suggested in [8], is as follows.

Fix a sufficiently small \( \eta \); this determines the back-off from extremality that the BST will achieve. This value, as shown in Appendix A, also needs to be small enough to ensure fast polarization of this stage. Choose \( L_0 \) and \( M_0 \), the BST parameters, and the number of BST levels \( n \) to ensure that a BST of length \( N = N(n) \) is \((\eta, \mathcal{L}, \mathcal{H})\)-monopolizing for the family of s/o-processes the codes will be used for, see Theorem 18 in Section V.C.

Further increase \( M_0 \), if necessary, to ensure that the fraction of medial s/o-pairs is as close to 1 as desired, see Lemma 2 in Section III-B.

After the slow polarization stage, all s/o-pairs in one of the sets \( \mathcal{L} \) or \( \mathcal{H} \) will be almost extremal. Suppose that we are in a channel-coding application. In this case, we need to ensure that the conditional entropy of any s/o-pair in \( \mathcal{L} \) will be less than \( \eta \).

This can only happen when the BST is used for channel-coding over a subfamily of s/o-processes whose conditional entropy rate is less than \(|\mathcal{L}|/(|\mathcal{L}| + |\mathcal{H}|)\).

Moreover, when \( M_0 \) is large enough, we obtain \(|\mathcal{L}| + |\mathcal{H}| \approx N \).

Now, take \( \hat{N} = 2^\eta \) copies of the BST of length \( N \). Apply multiple copies of Arıkan’s polar transform of length \( \hat{N} \), one for each medial s/o-pair in \( \mathcal{L} \) or \( \mathcal{H} \). Continuing our channel-coding example, take the first medial s/o-pair from \( \mathcal{L} \) of each of the BSTs and apply a length-\( \hat{N} \) Arıkan transform to them. Then, apply an Arıkan transform to the set of second medial s/o-pairs from \( \mathcal{L} \) of each of the BSTs, and so forth. The \( j \)th Arıkan transform operates only on the \( j \)th medial s/o-pair from \( \mathcal{L} \) from each BST. All other s/o-pairs are frozen and do not participate in the fast polarization stage. The fast stage operation is illustrated in Figure 6.

---

1If a universal code of rate 1/2 is required, then \( N(n) = 2^n(M_0 + 2L_0) \).
For other sizes of \( \mathcal{L} \) and \( \mathcal{H} \), see the strategies outlined in the discussion at the end of the previous section; these may entail a different kernel or combining multiple BSTs, and result in different BST lengths.
The BST ensures a universal bound on the conditional entropies of the s/o-pairs participating in the Arıkan transform. Crucially, this implies universal bounds on other distribution parameters: the Bhattacharyya parameter and total variation distance, see [14, Section III]. We design universal polar codes based on these universal bounds.

Continuing the channel-coding example, one can use the evolution of the Bhattacharyya parameter via polarization transforms to design polar codes for the fast stage to be used over a subfamily s/o-processes. Concretely, since Arıkan’s polar transform (see Example 1), is recursively defined, one may recursively compute upper bounds on the Bhattacharyya parameter of a transformed s/o-pair. Namely, \( Z_{n+1} \leq \kappa Z_n^M \), where \( \kappa > 1 \) is an easily-calculable universal parameter over the subfamily and the \( b_i \in \{1, 2\} \) are defined by the sequence of polarization transforms. For details, see [2, Proposition 5] and [13, Theorem 2]. In Appendix A we show that if \( \eta \) is small enough and \( \hat{N} \) is large enough, this sequence of upper bounds is ensured to polarize fast universally.

When \( M_0 \) is large and \( \eta \) is close to 0, there is negligible rate loss in both the slow and the fast polarization stages. In a universal channel-coding application, the input distribution is fixed and known in advance to the encoder and decoder. Thus, when designing the fast-stage polar codes, one would employ a Honda-Yamamoto [21] scheme.

In each fast-stage polar code, when \( \eta \) is small enough, almost all indices polarize fast over the subfamily. Continuing our channel-coding example, almost all transformed symbols of a polar code of length \( \hat{N} \) will have a very low Bhattacharyya parameter when conditioned on previous transformed symbols and the observation sequence. Achieving a desired input distribution requires identifying which transformed symbols also have very low total variation distance when conditioned

\[ j = (\ell - 1)N_0 + k, \quad 1 \leq \ell \leq 2^n, \quad 1 \leq k \leq N_0. \] (19)

\footnote{Similar relationships for the total variation distance also hold, see [14, Proposition 4 and Proposition 12].}
We call $\ell$ the b-block number and $k$ the b-index that correspond to index $j$.

Recall from Definition 6 that each medial level-$n$ symbol has $2^n$ medial level-0 indices as its base-ancestors. These base-ancestors are a subvector of $X_i^N$. Each of these level-0 indices has a different b-block number, computed via (19). We collect the sorted indices of these symbols in a vector as follows. From this point onwards, we use the term ‘ancestor’ to apply to both the symbol and its index; it will be clear from the context if we refer to the symbol or to its index.

Definition 8 (Base-vector and modulo-base-vector). The base-vector $b$ of a medial index $i$ is a row vector whose $t$th entry is the base-ancestor of $i$ from b-block $\ell$. Therefore,

$$(b)_\ell = (\ell - 1)N_0 + k$$

for some $L_0 + 1 \leq k \leq N_0 - L_0$.

The modulo-base-vector $\bar{b}$ of $i$ is defined by

$$(\bar{b})_\ell = (b)_\ell - (\ell - 1)N_0, \quad 1 \leq \ell \leq 2^n,$$

where $n$ is the level of $i$. This vector contains in its $\ell$th entry the b-index of $i$’s base-ancestor in the $\ell$th b-block. That is, $k$ in (20).

Remark 3. We only define base-vectors for medial indices. While it is possible to extend the definition to apply to lateral indices, this will not be of interest to us. This is afforded because the ancestors of medial indices can only be medial indices, so we will not need to consider lateral indices. In particular, equation (22) below is well-defined because each vector on the right-hand side is a modulo-base-vector of a base-vector. That is, in the memoryless case

$$P(F_i = 0 \mid F_1^{i-1}, Y_1^N) = P(F_i = 0 \mid F_1^{i-1}, Y_b),$$

where $Y_b = \{Y_{b_0}, Y_{b_1}, \ldots, Y_{b_2^n}\}$. We emphasize that the aforementioned assumption of a memoryless process was made solely for the purpose of motivating the base-vector. In fact, the base-vector is a product of the BST itself, and has nothing to do with the s/o-process being transformed. Henceforth, in this section we look at a BST as a transformation between two vectors, and study some of its properties.

To compute the base-vector of an index, we first compute its modulo-base-vector, and then use (21). The modulo-base-vectors are constructed recursively. To this end, we augment the notation for base- and modulo-base-vectors with the index and level specification. Thus, for $i \in [\text{med}(n)]$, we use $b_i^{(n)}$ and $\bar{b}_i^{(n)}$ to denote the base-vector and modulo-base-vector, respectively.

For a level-0 BST, the modulo-base-vector for medial index $L_0 + 1 \leq i \leq N_0 - L_0$ contains just one index:

$$\bar{b}_i^{(0)} = [i].$$

For higher levels, by Definition 6, the modulo-base-vectors are constructed by

$$\bar{b}_i^{(n+1)} = [\bar{b}_j^{(n)} \bar{b}_j^{(n)}], \quad j = \left\lfloor \frac{i}{2} \right\rfloor.$$  \hspace{1cm} (22)

Recall from Remark 1 that if $i \in [\text{med}_n(n+1)]$, then $i$ is even, so $i$ and $i + 1$ share the same base-vector.

Example 2. Consider a BST initialized with $L_0 = 3, M_0 = 6$. A level-0 BST is of length $N_1 = 2L_0 + M_0 = 12$. A level-1 BST is of length $N_1 = 2N_0 = 24$. The first medial index is $L_1 + 1 = (2L_0 + 1) + 1 = 8$. We have

$$\bar{b}_8^{(1)} = \bar{b}_9^{(1)} = [5 4], \quad \bar{b}_8^{(1)} = \bar{b}_9^{(1)} = [6 5],$$

and so on. A level-2 BST is of length $N_2 = 2N_1 = 48$, and its first medial index is $L_2 + 1 = (2L_1 + 1) + 1 = 16$. Thus,

$$\bar{b}_8^{(2)} = \bar{b}_9^{(2)} = [5 4 5 4], \quad \bar{b}_8^{(2)} = \bar{b}_9^{(2)} = [6 5 5 4].$$

A level-3 BST is of length $N_3 = 2N_2 = 96$, its first medial index is $L_3 + 1 = (2L_2 + 1) + 1 = 32$, and

$$\bar{b}_8^{(3)} = \bar{b}_9^{(3)} = [5 4 5 4 5 4 5 4], \quad \bar{b}_8^{(3)} = \bar{b}_9^{(3)} = [6 5 5 4 5 4 5 4].$$

Computing a base-vector, say $b_{30}^{(3)}$, is easily done using (21):

$$b_{30}^{(3)} = [6 17 29 40 53 64 77 88].$$

In Figure 7 we illustrate a portion of a level-3 BST and show the base-vector $b_{34}^{(3)} = b_{35}^{(3)}$.

Let $n \leq m$. Fix some $i \in [\text{med}(m)]$ and apply (22) recursively $m - n$ times. This expresses the modulo-base-vector of $i$ as a concatenation of $2^{m-n}$ level-$n$ modulo-base-vectors. These are the modulo-base-vectors of the level-$n$ ancestors of this level-$m$ index. In particular, the modulo-base-vector of any level-$n$ ancestor of $i$ is a sub-vector of $i$’s modulo-base-vector.

Example 2 (Continued). We can express the modulo-base-vector of level-3 index 34 as a concatenation of the modulo-base-vectors of its level-1 ancestors:

$$b_{34}^{(3)} = [6 5] [5 4] [5 4] [5 4].$$

Observe that in Example 2, the modulo-base-vectors of medial indices contain at least two and at most three distinct b-indices, and these b-indices are consecutive. This is not a coincidence, as the corollary to the following two lemmas will show.

Lemma 3. For any $i, i + 1 \in [\text{med}(n)]$ and any $1 \leq \ell \leq 2^n$ we have

$$b_{i+1}^{(n)} \geq b_i^{(n)}. \hspace{1cm} (23)$$

Proof: This follows from (22) by straightforward induction. Specifically, note that if the index $i$ on the left-hand-side of
where (22) increases, the indices with respect to its relevant-level BST (the rightmost are level-1). Let $i \in \mathbb{N}$, for some $i \in \mathbb{N}$, the formulas for computing the values of these elements. For any $i \in \mathbb{N}$, induction. First, we prove claim 1.

**Proof of Claim 1:** For $n = 0$ claim 1 is trivially true, as for any $i \in \mathbb{N}$, (22) holds for some $n \geq 0$; we will show it holds for $n + 1$ as well. Let $i \in \mathbb{N}(n + 1)$; by (22), $(b_i^{(n+1)})_1 = (b_i^{(n)})_1$, where $j = [i/2]$. Now, observe that for natural $i$,

$$i/2 = \begin{cases} i - 1/2 \\ i/2 \end{cases}.$$

Therefore,

$$(b_i^{(n+1)})_1 = (b_i^{(n)})_{[i/2]+1} = 1 + \left[ \frac{[i/2]}{2^n} \right] = 1 + \left[ \frac{[i-1/2]}{2^n} \right],$$

where (a) is by the induction assumption and (b) is by [31, equation 3.11].

**Proof of the left-hand side of (24):** For $n = 0$ and any $i \in \mathbb{N}(0)$, trivially $(b_i^{(0)})_1 = 1 + \lfloor (i - 1) \cdot 2^{-0} \rfloor = i$. Assume that the left-hand side of (24) holds for some $n \geq 0$; we will show it holds for $n + 1$ as well. Let $i \in \mathbb{N}(n + 1)$; by (22), $(b_i^{(n+1)})_{2n} = (b_i^{(n)})_{2n}$, where $j = [i/2]$. Therefore,

$$(b_i^{(n+1)})_{2n} = (b_i^{(n)})_{[i/2]} = \begin{cases} i/2 \\ i/2 \end{cases} = \begin{cases} i/2 \\ i/2 \end{cases} = \begin{cases} i/2 \\ i/2 \end{cases} = \begin{cases} i/2 \\ i/2 \end{cases},$$

where (a) is by the induction assumption and (b) is by [31, equation 3.11].

**Corollary 5.** If $n \geq 1$ then for any $i \in \mathbb{N}(n)$,

$$1 \leq \max_{\ell} (b_i^{(n)})_{\ell} - \min_{\ell} (b_i^{(n)})_{\ell} \leq 2.$$

**Proof:** This is an immediate consequence of Lemma 4. Specifically, if $[i/2^n] = r$ then

$$r \leq \frac{i}{2^n} < r + 1 \Rightarrow r - \frac{1}{2^n} \leq \frac{i - 1}{2^n} < r + 1 - \frac{1}{2^n}.$$ 

The ceiling operation $\lceil \cdot \rceil$ is monotonically increasing. Thus, we apply it to the three terms on the right-hand side to yield $r \leq \lfloor (i - 1)/2^n \rfloor \leq r + 1$.

**B. The Observation-Truncated BST**

The Observation-Truncated BST (OT-BST in short) is a variation on the BST that will be useful for analysis. It is defined recursively, just like the BST, but with a different initialization.

The BST may be looked at as a recursively-defined sequence of functions. Let $F_{X_Y} : G_{X_Y} \to B$ be the output of a level-$n$ BST with parameters $L_0$ and $M_0$ of s/o-block $X^{N_0} \to Y^{N_0}$. Recall that $X_1 \in X = \{0, 1\}$ and $Y_1 \in Y$ for any $i$, where $\overline{Y}$ is some finite alphabet. For any $i \in [\mathbb{N}(n)]$ there exist functions

$$f_{n,i} : X^{N_0} \to X,$$

$$g_{n,i} : X^{N_0} \times Y^{N_0} \to X^{i-1} \times Y^{N_0},$$

such that $f_{n,i}(X_1^{N_0}) = F_i$ and $g_{n,i}(X_1^{N_0}, Y_1^{N_0}) = G_i$. From (5), (10), and (13), they are recursively defined as follows. Initialization for any $i \in [\mathbb{N}(0)]$:

$$f_{0,i}(X_1^{N_0}) = X_i,$$

$$g_{0,i}(X_1^{N_0}, Y_1^{N_0}) = (X_1^{i-1}, Y_1^{N_0}).$$

(25a)

(25b)
Recursion for \( f_{n+1,i} \) for any \( i \in [\text{med}(n + 1)] \):

\[
\begin{align*}
   f_{n+1,i}(X_1^{N_{n+1}}) &= \begin{cases} 
   f_{n,j+1}(X_1^{N_n}) + f_{n,j}(X_2^{N_{n+1}^j}), & i = 2j, \\
   f_{n,j}(X_{2n+1}^{N_{n+1}^j}), & i = 2j + 1, 
   \end{cases} \\
   \text{where } &j \in [\text{med}(n)], \\
   &j \in [\text{med}(n)]. 
\end{align*}
\]

(26)

Recursion for \( g_{n+1,i} \) for any \( i \in [\text{med}(n + 1)] \):

\[
\begin{align*}
   g_{n+1,i}(X_1^{N_{n+1}}, Y_1^{N_{n+1}}) &= \begin{cases} 
   g_{n,j}(X_1^{2N_n}, Y_2^{N_{n+1}}), g_{n,j+1}(X_1^{N_n}, Y_1^{N_n}), & i = 2j, \\
   g_{n,j+1}(X_1^{N_n}, Y_1^{N_n}), g_{n,j}(X_2^{2N_{n+1}^j}, Y_2^{N_{n+1}^j}), & i = 2j + 1, 
   \end{cases} \\
   \text{where } &j \in [\text{med}(n)], \\
   &j \in [\text{med}(n)]. 
\end{align*}
\]

(27)

In the recursion for \( g_{n+1,i}(X_1^{N_{n+1}}, Y_1^{N_{n+1}}) \) where \( i = 2j \) we differentiate between the cases \( j \in [\text{med}(n)] \) and \( j \in [\text{med}(n)] \) to ensure that, for even \( i \), the first part of the observation is an observation from \([\text{med}(n)]\) and the second part is an observation from \([\text{med}(n)]\). This is an artifact of the mental indices alternating between blocks, see Figure 5. This subtlety will be important for a technicality in the proof of Lemma 11 below. For all other purposes, the reader is encouraged to disregard this rather technical distinction.

We concentrate here only on mental indices, because our analysis will focus on mental indices. The recursion (26), (27) is well-defined, as mental indices are only ever generated from mental indices (see Remark 3), so nowhere in the recursion will a non-mental index appear.

The observation-truncated BST is also a recursively-defined sequence of functions \( \bar{f}_{n,i} \) and \( \bar{g}_{n,i} \). The recursion for these functions is given by (26) and (27) and is governed by the same two parameters, \( L_0 \) and \( M_0 \), as the BST. However, the OT-BST has a different initialization than that of the BST. The initialization for the OT-BST is, for any \( i \in [\text{med}(0)] \),

\[
\begin{align*}
   \bar{f}_{0,i}(X_1^{N_0}) &= X_i, \\
   \bar{g}_{0,i}(X_1^{N_0}, Y_1^{N_0}) &= (X_i - L_0, Y_i + L_0).
\end{align*}
\]

(28a)

(28b)

By comparing (25) and (28), two observations are made. First, \( \bar{f}_{n,i} = f_{n,i} \) for any \( i \in [\text{med}(n)] \). Second, there exists a mapping \( \gamma_{n,i} \) from \( g_{n,i} \) to \( \bar{g}_{n,i} \). That is, given \( G_i = g_{n,i}(X_1^{N_n}, Y_1^{N_n}) \), one may compute

\[
\bar{g}_{n,i}(X_1^{N_n}, Y_1^{N_n}) = \gamma_{n,i}(g_{n,i}(X_1^{N_n}, Y_1^{N_n})) = \gamma_{n,i}(G_i).
\]

This is clear from the initialization step, and for the remaining steps it follows from the recursive definition (27) and since \( f_{n,i} = \bar{f}_{n,i} \).

The domains for \( f_{n,i}, \bar{f}_{n,i}, g_{n,i}, \bar{g}_{n,i} \) are over specified. Not all inputs of these functions are relevant. The relevant domain of these functions may be expressed using the base-vector of \( i \). To this end, we recall the following notation. For any vector of indices \( i = [i_1, i_2, \ldots, i_k] \), natural numbers \( L, M \), and a sequence of random variables \( X_j \), we denote

\[
\begin{align*}
   X_i &= (X_{i_1}, X_{i_2}, \ldots, X_{i_k}), \\
   X_{i-L} &= (X_{i_1-L}, X_{i_2-L}, \ldots, X_{i_k-L}), \\
   X_{i+M} &= (X_{i_1+M}, X_{i_2+M}, \ldots, X_{i_k+M}).
\end{align*}
\]

(29a)

(29b)

(29c)

Now, let \( b \) be the base-vector of level-\( n \) index \( i \). Then, \( f_{n,i} \) and \( \bar{f}_{n,i} \) are actually functions of \( X_b \). This follows from the recursive definitions of the functions and the base-vector. With some abuse of notation we henceforth write

\[
f_{n,i}(X_b) = f_{n,i}(X_b).
\]

Similarly, by (25b), (27), and (28b),

\[
\begin{align*}
   g_{n,i}(X_b^b Y_a) &= g_{n,i}(X_b^b Y_a^a), \\
   \bar{g}_{n,i}(X_b^b Y_a) &= \bar{g}_{n,i}(X_b^b Y_a^a),
\end{align*}
\]

where we denoted

\[
a = [1 \ 2N_0 + 1 \ 2N_0 + 2 \cdots \ (2^n - 1)N_0 + 1]
\]

\[
z = [N_0 \ 2N_0 \ 3N_0 \ \cdots \ 2^n N_0].
\]

Note that \( Y_a^a = Y_1^{N_n} \).

Example 2 (Continued). For a level-3 BST initialized with \( L_0 = 3, M_0 = 6 \), consider \( f_{3,34} \) and \( f_{3,35} \). The base-vector for either index 34 or 35 is

\[
b = [6 \ 17 \ 29 \ 40 \ 53 \ 64 \ 77 \ 88].
\]

We have (see Figure 7):

\[
F_{34} = f_{3,34}(X_b) = X_6 + X_{17} + X_{40} + X_{77} + X_{88},
\]

\[
F_{35} = f_{3,35}(X_b) = X_6 + X_{17} + X_{40}.
\]

Recall that \( b \) is the base-vector of level-\( n \) index \( i \). From the recursive definition (27), we observe that we can compute \( X_{b-L_0}^b \) from \( g_{n,i}(X_b^b Y_a^b) \). This is easily shown by induction. It is trivially true for \( n = 0 \). Assume that this holds for \( n \geq 0 \) for any mental index; we will show it holds for \( n + 1 \) as well. Indeed, write \( b = [b_1, b_2] \), where \( b_1 \) and \( b_2 \) are of length \( 2n+1 \). By the recursive definition of \( b \), (22), the recursion (27) becomes

\[
\begin{align*}
   \bar{g}_{n+1,i}(X_b^b Y_a^b) &= \bar{g}_{n+1,i}(X_b^b Y_a^b) \in [\text{med}(n)], \\
   \bar{g}_{n+1,i}(X_b^b Y_a^b) &= \bar{g}_{n+1,i}(X_b^b Y_a^b) \in [\text{med}(n)],
\end{align*}
\]

By the induction hypothesis, we can compute \( X_{b-L_0}^b \) from \( g_{n+1,i}(X_b^b Y_a^b) \) and \( y_{b-L_0}^b \) from \( g_{n,i}(X_b^b Y_a^b) \). In other words, we can compute \( y_{b-L_0}^b \) from \( g_{n+1,i}(X_b^b Y_a^b) \). Of course, one can also compute \( y_{b-L_0}^b \) from \( g_{n,i}(X_b^b Y_a^b) \). Therefore, recalling that \( = \) between two vectors means that there is a one-to-one mapping between either one and the other that is independent of either vector,

\[
\bar{g}_{n+1,i}(X_b^b Y_a^b Y_{b-L_0}^b) \in \bar{g}_{n+1,i}(X_b^b Y_{a-L_0}^b Y_{b-L_0}^b) \in [\text{med}(n)].
\]

(30)
We saw above that given $G_i = g_{n,i}(X^{b}_{n}, Y^{b}_{n})$ one can compute $\tilde{G}_i = \tilde{g}_{n,i}(X^{b}_{n-L_0}, Y^{b}_{n-L_0})$. In fact, more is true. We can compute from $G_i$ two quantities: $\tilde{G}_i$, which is a function of $(X^{b}_{n-L_0}, Y^{b}_{n-L_0})$, and $\hat{G}_i$, which consists of $(X^{b}_{n-L_0-1}, Y^{b}_{n-L_0-1}, Y^{b}_{n-L_0+1})$. Thus, we may write

$$G_i = g_{n,i}(X^{b}_{n}, Y^{b}_{n}) = (\tilde{G}_i, \hat{G}_i),$$

where

$$\tilde{G}_i = \tilde{g}_{n,i}(X^{b}_{n-L_0}, Y^{b}_{n-L_0}),$$

$$\hat{G}_i = (X^{b}_{n-L_0-1}, Y^{b}_{n-L_0-1}, Y^{b}_{n-L_0+1}).$$

This follows by induction similar to the one above. Indeed, this is obvious for the initialization step by comparing (25b) and (28b), and the induction step follows, as above, from the recursive definition of the base-vector (22) and from (27).

**Remark 4.** At this point, the reader may be wondering why we used the notation $\tilde{G}_i, \hat{G}_i$ rather than $\tilde{G}_i, \hat{G}_i$. The reason is that we reserve the latter notation to the result of the OT-BST when applied for a different process, the block-independent process, that we introduce in Section V-B. The notation for the block-independent process will use tildes. Our main use of the OT-BST will be for the block-independent process.

We conclude this section with a note on terminology. The OT-BST is not an H-transform. That said, we borrow some terminology from H-transforms and apply it to the OT-BST. Specifically, for level-$n$ index $i$ with base-vector $b$ we call $f_{n,i}(X_b)$ an OT-transformed index. The conditional entropy of OT-transformed level-$n$ index $i$ is $H(f_{n,i}(X_b)|\tilde{g}_{n,i}(X^{b}_{n-L_0}, Y^{b}_{n-L_0}))$. Finally, for $\eta > 0$ and index sets $\mathcal{L}, \mathcal{H} \subseteq \{1, 2, \ldots, N_n\}$, the OT-BST is $(\eta, \mathcal{L}, \mathcal{H})$-monopolarizing if either $H(f_{n,i}(X_b)|\tilde{g}_{n,i}(X^{b}_{n-L_0}, Y^{b}_{n-L_0})) < \eta$ for all $i \in \mathcal{L}$, or $H(f_{n,i}(X_b)|\tilde{g}_{n,i}(X^{b}_{n-L_0}, Y^{b}_{n-L_0})) > 1 - \eta$ for all $i \in \mathcal{H}$.

**V. THE BST IS MONOPOLARIZING**

For a suitable family of s/o-processes, the BST is monopolarizing. We now describe this family and establish that the BST is monopolarizing for it.

**A. A Probabilistic Model with Memory**

The s/o-processes for which we prove that the BST is monopolarizing share a certain probabilistic structure. That is, the distribution of the s/o-process $X_s \rightarrow Y_s$ has a specific form: it depends on an underlying Markov sequence, $S_j, j \in \mathbb{Z}$.

We assume throughout that, for any $j$, $X_j$ is binary, $Y_j \in \mathcal{Y}$, and $S_j \in \mathcal{S}$, where $\mathcal{X}, \mathcal{Y}$ are finite alphabets.

**Definition 9 (FAIM process).** A strictly stationary process $(S_j, X_j, Y_j), j \in \mathbb{Z}$ is called a Finite-State, Aperiodic, Irreducible, Markov (FAIM) process if it has, for any $j$,

$$P_{S_j X_j Y_j} = P(S_{j-1} | S_{j-1}) \cdot P(S_j | S_{j-1}) \cdot P(X_j | S_j) \cdot P(Y_j | S_j),$$

and $S_j, j \in \mathbb{Z}$ is a finite-state, homogeneous, irreducible, and aperiodic stationary Markov chain.

An s/o-process $X_s \rightarrow Y_s$ whose joint distribution is derived from a FAIM process $(S_j, X_j, Y_j)$ is called a FAIM-derived s/o-process.

Equation (32) implies that conditioned on $S_{j-1}$, the random variables $S_j, X_j, Y_j$ are independent of $S_{j-1}, X_i, Y_i$, for any $l < j < k$. Furthermore, $X_j, Y_j$ are a function (possibly probabilistic) of $S_j$. FAIM processes are described in detail in [14].

**Remark 5.** The definition of FAIM processes in [14] did not include the rightmost equality of (32). However, by suitably redefining the state of the process (for example, take $(S_j, S_{j-1})$ as the state at time $j$), we may obtain the rightmost equality of (32) from its leftmost equality. Therefore, there is no loss of generality in the definition of a FAIM process given here as compared to the one in [14].

In the following lemma we prove an important property of FAIM processes. Informally, it implies that two s/o-blocks that are sufficiently far apart—that is, the last index of the first s/o-block is sufficiently less than the first index of the second s/o-block—are approximately independent.

**Lemma 6.** If $X_s \rightarrow Y_s$ is a FAIM-derived s/o-process, there exist sequences $\psi_k, \phi_k, k \geq 0$, such that for any $L \leq M \in \mathbb{Z}$,

$$P_{X_s^{L} Y_s^{L}, X_s^{M} Y_s^{M}} \leq \psi_{M-L} P_{X_s^{L} Y_s^{L}, X_s^{M} Y_s^{M}} + \phi_{M-L} P_{X_s^{L} Y_s^{L}, X_s^{M} Y_s^{M}}$$

This sequence $\psi_k$ is nonincreasing and the sequence $\phi_k$ is nondecreasing. Both $\psi_k$ and $\phi_k$ tend to 1 exponentially fast as $k \rightarrow \infty$.

The sequences $\psi_k$ and $\phi_k$ are called mixing sequences. Part of the lemma, namely (33a), was established in [14, Lemma 5], and the proof for (33b) is similar. For completeness, we provide a proof in Appendix B. We note at this point that for $k \geq 1$ we may take

$$\psi_k = \max_{s, \sigma} \frac{P(S_k = s, S_{k-1} = \sigma)}{P(S_k = \sigma)}$$

in (33). These are well-defined because the Markov chain $S_j, j \in \mathbb{Z}$ is finite-state, irreducible, and aperiodic. As a result, its stationary distribution is positive: $P(S_k = s) > 0$ for any $s \in \mathcal{S}$ and $k \in \mathbb{Z}$, [32, Theorem 4.2].

It is immediately evident that for any $k \geq 1, 0 \leq \psi_k < \infty$ and $0 \leq \phi_k \leq 1$. It is possible, however, that for small values of $k$, we will have $\phi_k = 0$. Nevertheless, Lemma 6 ensures that if $k$ is large enough, $\phi_k$ will be positive; in fact, by increasing $k$ it can be as close to 1 as desired.

Lemma 6 ensures that s/o-blocks of a FAIM-derived process become almost independent when sufficiently far apart. We will need a separate property that explores what happens when a single s/o-block of a FAIM process is large enough. Specifically, we will be interested in FAIM processes that, in a sense, “forget”
their past. In a forgetful FAIM process, the initial and final states of a sufficiently large s/o-block are almost independent both when given just the observations or when given the symbols and observations jointly. A precise definition of a forgetful FAIM process follows.

**Definition 10** (Forgetful FAIM process). A FAIM process \((S_j, X_j, Y_j), j \in \mathbb{Z}\) is said to be forgetful if for any \(\epsilon > 0\) there exists a natural number \(\lambda\) such that if \(k \geq \lambda\) then

\[
I(S_j; S_k | X_j^k, Y_j^k) \leq \epsilon, \quad (34a)
\]
\[
I(S_j; S_k | Y_j^k) \leq \epsilon. \quad (34b)
\]

We call \(\epsilon\) the forgetfulness of the s/o-process, and \(\lambda\) the recollection of the process. The recollection for a given \(\epsilon\) is called '\(\epsilon\)-recollection.' The forgetfulness for a given \(\lambda\) is called '\(\lambda\)-forgetfulness.'

We say that FAIM-derived s/o-process \(X_* \rightarrow Y_*\) is forgetful if it is derived from a forgetful FAIM process.

Several remarks are in order.

1) A sufficient condition for a FAIM process to be forgetful (Condition K), as well as how to compute the recollection for a given \(\epsilon\), are detailed in Section VIII (see also Example 7 in that section). In particular, forgetful FAIM processes do exist. For processes that satisfy Condition K, the forgetfulness decreases exponentially with the recollection.

2) Somewhat unintuitively, a FAIM process need not to be forgetful. See Example 3 below for an example of a FAIM process that is not forgetful.

3) Both conditions (34a) and (34b) are required: neither condition implies the other. We demonstrate this unintuitive fact in Example 4 below.

4) Equations (34a) and (34b) imply, by the data processing inequality (2) and the Markov property (32), that for any \(k \geq \lambda, \ell \leq 1,\) and \(m \geq k,\)

\[
I(S_j; S_m | X_j^k, Y_j^\ell) \leq \epsilon, \quad (35a)
\]
\[
I(S_j; S_m | Y_j^\ell) \leq \epsilon. \quad (35b)
\]

**Example 3.** This example is due to [19, Section 10]. In Figure 8 we illustrate the process \((S_j, Y_j)\). Specifically, the Markov chain \(S_j\) has transition matrix

\[
M = \begin{bmatrix}
0 & 1/2 & 0 & 1/2 \\
1/2 & 0 & 1/2 & 0 \\
1/2 & 0 & 0 & 1/2 \\
0 & 1/2 & 1/2 & 0
\end{bmatrix},
\]

and the observation \(Y_j\) is given by

\[
Y_j = \begin{cases}
\text{a, if } S_j \in \{1, 2\}, \\
\text{b, if } S_j \in \{3, 4\}.
\end{cases} \quad (36)
\]

In this example we will not be interested in \(X_j\). This is a FAIM process: the Markov chain \(S_j\) is finite-state, aperiodic, and irreducible; indeed, \(M^3 > 0\).

From the observation \(Y_j\) we can infer whether state \(S_j\) is in the top half or the bottom half of Figure 8. For two consecutive observations to differ, the process must transition from a state in one half of Figure 8 to the other. Given a sequence of observations, our best guess for the next state is equi-probable among two states. For example, given the observation sequence \(Y_1 = a, Y_2 = b, \ldots, Y_k = b\), we know that \(S_k \in \{3, 4\}\), but \(S_k\) could be either 3 or 4 with equal probability.

Assume now that, in addition to the observation sequence, we are told the state at time 1. Say, \(S_1 = 1\) (accordingly, \(Y_1 = a\)). The observations are tied to transitions from one half of Figure 8 to the other half, so that one can trace the state: \(Y_2 = a\) implies that \(S_2 = 1\). Then, \(Y_3 = b\) implies that \(S_3 = 3\), and so on. In this manner, we are able to find \(S_k\) precisely.

We have demonstrated that in this example, \(I(S_j; S_k | Y_j^k)\) cannot vanish with \(k\), so this process is not forgetful.

**Example 4.** Let \(S_j\) be as in Example 3. We now construct two FAIM processes. For the first process, \(I(S_j; S_k | X_j^k, Y_j^k)\) will vanish with \(k\) but \(I(S_j; S_k | Y_j^k)\) will not. For the second process, \(I(S_j; S_k | X_j^k, Y_j^k)\) will not vanish with \(k\) but \(I(S_j; S_k | Y_j^k)\) will.

- Let \(X_j = S_j\) and \(Y_j\) as in (36). Then, \(I(S_j; S_k | X_j^k, Y_j^k) = I(S_j; S_k | S_j^k) = 0\) trivially. On the other hand, as shown in Example 3, \(I(S_j; S_k | Y_j^k)\) does not vanish for any \(k\).
- Let \(X_j\) be given by (36) (that is, \(X_j = a\) if \(S_j \in \{1, 2\}\) and \(X_j = b\) otherwise) and \(Y_j = 0\). Then, \(I(S_j; S_k | X_j^k, Y_j^k)\) cannot vanish with \(k\), as shown in Example 3. On the other hand, \(I(S_j; S_k | Y_j^k) = I(S_j; S_k) \rightarrow 0\), since the Markov chain \(S_j\) is finite-state, aperiodic, and irreducible (see, e.g., [32, Theorem 4.3]).

Assume we have a forgetful FAIM process, and we apply it to a level-0 BST, initialized with \(L_0\) that is greater than its \(\epsilon\)-recollection. We expect that in this case, all medial s/o-pairs will have approximately the same conditional entropy. This is indeed the case, as we will soon show in Lemma 9. Moreover, we will see in Corollary 10 that this conditional entropy cannot veer much from the conditional entropy rate of the s/o-process.

First, however, we require an additional lemma.

**Lemma 7.** Let \((S_j, X_j, Y_j)\) be a forgetful FAIM process. Then, for every \(\epsilon > 0\) there exists a natural number \(\lambda\) such that for any integers \(m, \ell, k\) such that \(|m, \ell, k| \geq k \geq \lambda\) we have

\[
I(S_j; S_{-k}, S_k | X_j^m, Y_j^n) \leq 2\epsilon. \quad (37)
\]
This is a consequence of (34). To prove it, we take \( \lambda \) as the \( \epsilon \)-recollect of the process, and make multiple uses of (2), which are possible due to the Markov property (32). A detailed proof can be found in Appendix B.

Lemma 7 shows that the mutual information between a state and two surrounding states vanishes when given a sequence of observations between the surrounding states. The following corollary shows that this is also the case when considering the mutual information between a sequence of states and a sequence of surrounding states. This will be useful in the sequel.

**Corollary 8.** Let \( (S_j, X_j, Y_j) \) be a forgetful FAIM process. Then, for every \( \epsilon > 0 \) there exists a natural number \( \lambda \) such that for any positive natural numbers \( k, i_1, i_2, \ldots, i_k \), and \( L_0 \) that satisfy \( L_0 \geq \lambda \) and

\[
i_1 - L_0 \leq i_1 \leq i_1 + L_0 \leq i_2 - L_0 \leq i_2 \leq \cdots \leq i_k \leq i_k + L_0
\]

we have

\[
I(S_i; S_{i-L_0}, S_{i+L_0} | X^{i-1}_{L_0}, Y^{i+L_0}_{L_0}) \leq k \cdot 2\epsilon,
\]

where

\[
i = [i_1 \ i_2 \ \cdots \ i_k].
\]

In the statement of the corollary, we used the notation of (29). The proof of the corollary is relegated to Appendix B.

In the next lemma, we show that, for a forgetful FAIM-derived s/o-process, all medial s/o-pairs in a level-0 BST have approximately the same conditional entropy,

\[
\tilde{\mathcal{R}} \triangleq H(X_i | X^{i-1}_{L_0}, Y^{i+L_0}_{L_0}).
\]

By stationarity, \( \tilde{\mathcal{R}} \) is indeed independent of \( i \).

**Lemma 9.** Let \( X_\ast \rightarrow Y_\ast \) be a forgetful FAIM-derived s/o-process with \( \epsilon \)-recollect \( \lambda \). Let \( L_0 \geq \lambda \) and \( M_0 \geq 1 \), and denote \( N_0 = 2L_0 + M_0 \). Then, for any \( L_0 + 1 = i \leq L_0 + M_0 \) we have

\[
0 \leq \tilde{\mathcal{R}} - H(X_i | X^{i-1}_{L_0}, Y^{N_0}_{L_0}) \leq 2\epsilon.
\]

**Proof:** Observe that

\[
\begin{align*}
\tilde{\mathcal{R}} - H(X_i | X^{i-1}_{L_0}, Y^{N_0}_{L_0}) & = H(X_i | X^{i-1}_{L_0}, Y^{i+L_0}_{L_0}) - H(X_i | X^{i-1}_{L_0}, Y^{N_0}_{L_0}) \\
& = I(X_i; X^{i-1}_{L_0}, Y^{i+L_0}_{L_0}) + I(X_i; X^{i-1}_{L_0}, Y^{N_0}_{L_0}) - I(X_i; X^{i-1}_{L_0}, Y^{i+L_0}_{L_0})
\end{align*}
\]

This right-hand side is nonnegative. It remains to upper-bound it by \( 2\epsilon \) to establish (39).

By (32) and the data processing inequality (2) used twice, we obtain

\[
2\epsilon \geq I \left( S_i; S_{i-L_0}, S_{i+L_0} | X^{i-1}_{i-L_0}, Y^{i+L_0}_{i-L_0} \right)
\]

\[
= I \left( X_i; S_{i-L_0}, S_{i+L_0} | X^{i-1}_{i-L_0}, Y^{i+L_0}_{i-L_0} \right)
\]

\[
\geq I \left( X_i; X^{i-1}_{i-L_0}, Y^{i+L_0}_{i-L_0} \right)
\]

\[
\geq I \left( X_i; X^{i-1}_{i-L_0}, Y^{i+L_0}_{i-L_0} \right).
\]

We now detail the Markov chains used for the inequalities, both using (32). Inequality (a) is due to \( (S_i, X_i, Y_i) \rightarrow (S_{i-L_0}, S_{i+L_0}) \rightarrow (X_i, Y_i) \).

By (32) and the data processing inequality (2) used twice, we obtain

\[
2\epsilon \geq I \left( S_i; S_{i-L_0}, S_{i+L_0} | X^{i-1}_{i-L_0}, Y^{i+L_0}_{i-L_0} \right)
\]

\[
= I \left( X_i; S_{i-L_0}, S_{i+L_0} | X^{i-1}_{i-L_0}, Y^{i+L_0}_{i-L_0} \right)
\]

\[
= I \left( X_i; X^{i-1}_{i-L_0}, Y^{i+L_0}_{i-L_0} \right)
\]

This completes the proof.

The following corollary shows that, for a forgetful FAIM-derived s/o-process, \( \tilde{\mathcal{R}} \) is approximately equal to the conditional entropy rate of the s/o-process.

**Corollary 10.** Under the same setting as Lemma 9,

\[
|\mathcal{R}(X_* | Y_*) - \tilde{\mathcal{R}}| \leq 2\epsilon,
\]

**Proof:** For any \( \xi > 0 \), let \( N = N(\xi) \geq 2L_0 \) be large enough so that \( |\mathcal{R}(X_* | Y_*) - H(X^N_1 | Y^N_1) / N| \leq \xi / 2 \) and \( 2L_0 / N \leq \xi / 2 \). Then,

\[
|\mathcal{R}(X_* | Y_*) - \tilde{\mathcal{R}}| \leq \xi / 2 + 1 / N \sum_{i=1}^{N} |H(X_i | X^{i-1}_1, Y^{N_0}_1) - \tilde{\mathcal{R}}| \leq \xi / 2 + 1 / N \sum_{i=L_0+1}^{N} |H(X_i | X^{i-1}_1, Y^{N_0}_1) - \tilde{\mathcal{R}}| \leq \xi / 2 + 2L_0 / N \leq 2\xi / N \leq 2\epsilon + \xi,
\]

where (a) and (b) are by the triangle inequality; (c) is because \( |H(X_i | X^{i-1}_1, Y^{N_0}_1) - \tilde{\mathcal{R}}| \leq \max(\tilde{\mathcal{R}}, H(X_i | X^{i-1}_1, Y^{N_0}_1)) \leq 1 \), where the latter inequality holds since \( X_i \) is binary; finally, (d) is by Lemma 9, with \( N_0 \) replaced with \( N \). The above holds for any \( \xi > 0 \), so it holds for \( \xi = 0 \) as well.

**B. The Block-Independent Process**

We will prove in Section V-C that the BST is monopolarizing with the help of another process, the block-independent process, that we now introduce. We will show that an OT-BST is monopolarizing when applied to the block-independent process. It turns out that the result of an OT-BST applied to the block-independent process is approximately the same as the result of a BST applied to a forgetful FAIM-derived process, provided that the transform parameters are carefully chosen. Therefore, monopolarization of the OT-BST of the block-independent process will be of vital importance in proving that the BST is monopolarizing.
Let $N_0 = 2^n N_0$, where $N_0 = 2L_0 + M_0$. Denote by $P_{X_1^N,Y_1^N}$ the joint distribution of $(X_1^N, Y_1^N)$. By marginalizing $P_{X_1^N,Y_1^N}$, we obtain the distribution of a single b-block, $P_{X_1^N,Y_1^N}$, which, by stationarity, is independent of $t$.

**Definition 11** (Block-Independent Process). The block-independent process $(X_*) \rightarrow (Y_*)$ with parameter $N_0$, is distributed according to

$$ (X_1^N, Y_1^N) \sim \frac{2^n}{\ell - 1} \cdot P_{X_1^N,Y_1^N} \cdot Y_{X_1^N,Y_1^N}^{(\ell-1)} + 1, $$

That is, $b$-blocks of length $N_0$ are independent in this distribution.

If $b = [b_1, b_2, \ldots, b_{2^n}]$ is the base-vector of a level-$n$ medial index, we have

$$ X_1^{b(\ell-1)}_r \sim \frac{2^n}{\ell - 1} \cdot P_{X_1^{b(\ell-1)}_r,Y^{b(\ell-1)}_r} \cdot Y_{X_1^{b(\ell-1)}_r,Y^{b(\ell-1)}_r}^{(\ell-1)} + 1, $$

where $P_{X_1^{b(\ell-1)}_r,Y^{b(\ell-1)}_r}$ is obtained from $P_{X_1^N,Y_1^N}$ by marginalization. Note that each $b_\ell$ is medial, $(X_1^{b(\ell-1)}_r, Y^{b(\ell-1)}_r)$ is wholly contained in a $b$-block with $b$-block number $\ell$.

Throughout this section index $i \in [\text{med}(n)]$ has base-vector $b = [b_1, b_2, \ldots, b_{2^n}]$, and index $j \in [\text{med}(n)]$ has base-vector $d = [d_1, d_2, \ldots, d_{2^n}]$. We also denote

$$ a = [1, N_0 + 1, 2N_0 + 1, \ldots, (2^n - 1)N_0 + 1], $$

$$ z = [N_0, 2N_0, 3N_0, \ldots, 2^n N_0]. $$

Recalling the definitions of $\bar{f}_{n,i}$ and $\bar{g}_{n,i}$ at the beginning of Section IV-B, we define

$$ F_i = \bar{f}_{n,i}(X_{b}), \quad G_i = \bar{g}_{n,i}(X_{b}(X_{b}^{b(\ell-1)}_r, Y^{b(\ell-1)}_r)). $$

$$ F_i = \bar{f}_{n,i}(X_{d}), \quad G_i = \bar{g}_{n,i}(X_{d}(X_{d}^{b(\ell-1)}_r, Y^{b(\ell-1)}_r)). $$

The joint distribution of $(X_1^{b(\ell-1)}_r, Y^{b(\ell-1)}_r)$ is given by (42) with $b$ as the base-vector of $i$. The joint distribution of $(X_1^{d(\ell-1)}_r, Y^{d(\ell-1)}_r)$ is given by (42) with $b$ set to $d$, the base-vector of $j$.

Recall from (38) that we denoted $\mathcal{R} = H(X_1^{i(\ell-1)}_r, Y^{i(\ell-1)}_r)$, which, by stationarity, is independent of $i$. We wish to show that there exists $\delta_n > 0$, independent of $i$, such that if $i \in [\text{med}(n)]$ then $H(F_i) \sim \mathcal{R} \sim \delta_n$ and if $i \in [\text{med}(n)]$ then $H(F_i) \sim \mathcal{R} \sim \delta_n$. This will follow as a corollary to the following lemma.

**Lemma 11.** Suppose that either $i, j \in [\text{med}(n)]$ or $i, j \in [\text{med}(n)]$. Then, the joint distribution of $(F_i, G_i)$ is the same as the joint distribution of $(\bar{F}_j, \bar{G}_j)$.

**Proof:** We use induction by stationarity and the initialization of the OT-BST (28). Indeed, in this case, $F_i = \bar{X}_i, F_j = \bar{X}_j, G_i = (X_1^{i(\ell-1)}_r, Y^{i(\ell-1)}_r),$ and $G_j = (X_1^{j(\ell-1)}_r, Y^{j(\ell-1)}_r)$. Stationarity implies that the joint distribution of $(F_i, G_i)$ is the same as the joint distribution of $(\bar{F}_j, \bar{G}_j)$.

Assume the claim is true for some $n - 1 \geq 0$. We now show it holds for $n$.

Denote $i' = [i/2]$ and $j' = [j/2]$. We write $b = [b_1, b_2]$ and $d = [d_1, d_2]$, where $b_1, b_2, d_1, d_2$ are vectors of length $2^{n-1}$. Then, $b_1$ is the base-vector of $i' + 1$, and $b_2$ is the base-vector of $i'$, see (22). Similarly, $d_1$ is the base-vector of $j' + 1$, and $d_2$ is the base-vector of $j'$. Denote

$$ U_{i',i'} = f_{n-1,i',i'}(X_{b_1}), \quad U_{j',j'} = f_{n-1,j',j'}(X_{b_2}), $$

$$ V_{i'} = f_{n-1,i'}(X_{b_1}), \quad V_{j'} = f_{n-1,j'}(X_{b_2}). $$

Of the two s/o-pairs $U_{i',i'} \rightarrow U_{j',j'}$ and $V_{i'} \rightarrow V_{j'}$, one is in $[\text{med}(n-1)]$ and the other in $[\text{med}(n-1)]$. We denote by $\mathcal{T}_i$ the pair that is in $[\text{med}(n-1)]$ and by $\mathcal{T}_i^{+}$ the pair that is in $[\text{med}(n-1)]$. That is,

$$ \mathcal{T}_i = \{(V_{i'}, V_{i'}), i' \in [\text{med}(n-1)]\}, $$

$$ \mathcal{T}_i^{+} = \{(U_{i',i'}, U_{i',i'}'), i' \in [\text{med}(n-1)]\}. $$

We similarly define $U_{j',j'}$, $V_{j'}$, $Q_{j',j'}$, $R_{j'}$, $T_j$, and $T_j^{+}$ (with $b$ replaced with $d$ and $i'$ replaced with $j'$).

For the BI-process, $b$-blocks are independent. In particular, by (42), $(X_{b_1}^{b(\ell-1)}_r, Y^{b(\ell-1)}_r)$ is independent of $(X_{b_2}^{b(\ell-1)}_r, Y^{b(\ell-1)}_r)$. Hence, $\mathcal{T}_i$ and $\mathcal{T}_i^{+}$ are independent. Similarly, $\mathcal{T}_j$ and $\mathcal{T}_j^{+}$ are independent. By the induction hypothesis, $\mathcal{T}_i$ and $\mathcal{T}_j$ have the same distribution; $\mathcal{T}_i^{+}$ and $\mathcal{T}_j^{+}$ are also equi-distributed. By block-independence, the joint distribution of $(\mathcal{T}_i, \mathcal{T}_j^{+})$ is the same as the joint distribution of $(\mathcal{T}_i, \mathcal{T}_j^{+})$.

Assume first that $i, j \in [\text{med}(n)]$. Then we have by (26) and (27),

$$ \bar{F}_i = \bar{U}_{i',i'} + \bar{V}_{i'}, \quad \bar{G}_i = \{(\bar{R}_{i'}, \bar{Q}_{i',i'}), i' \in [\text{med}(n-1)]\}, $$

and

$$ \bar{F}_j = \bar{U}_{j',j'} + \bar{V}_{j'}, \quad \bar{G}_j = \{(\bar{R}_{j'}, \bar{Q}_{j',j'}), j' \in [\text{med}(n-1)]\}. $$

(45) and (46), the mapping from $(\mathcal{T}_i, \mathcal{T}_j^{+})$ to $(\bar{F}_i, \bar{G}_i)$ is the same as the mapping from $(\mathcal{T}_i, \mathcal{T}_j^{+})$ to $(\bar{F}_j, \bar{G}_j)$. We conclude that the joint distribution of $(\bar{F}_i, \bar{G}_i)$ is the same as the joint distribution of $(\bar{F}_j, \bar{G}_j)$.

For the case where $i, j \in [\text{med}(n)]$, we have by (26),

$$ \bar{F}_i = \bar{U}_{i',i'} + \bar{V}_{i'}, \quad i' \in [\text{med}(n-1)]. $$

Observe that $\bar{F}_i$ is always a symbol in $[\text{med}(n-1)]$. Further recall from (26) that, since $i - 1 \in [\text{med}(n)]$, we have $\bar{F}_{i-1} = \bar{U}_{i',i'} + \bar{V}_{i'}$, so that $\bar{F}_i + \bar{F}_{i-1}$ is a symbol from $[\text{med}(n-1)]$.

By (26) and (27),

$$ (\bar{F}_i, \bar{G}_i) = (\bar{F}_i, \bar{F}_{i-1}, \bar{G}_{i-1}) \equiv (\bar{F}_i, \bar{F}_i + \bar{F}_{i-1}, \bar{G}_{i-1}). $$

(47)

Similarly,

$$ (\bar{F}_j, \bar{G}_j) = (\bar{F}_j, \bar{F}_{j-1}, \bar{G}_{j-1}) \equiv (\bar{F}_j, \bar{F}_j + \bar{F}_{j-1}, \bar{G}_{j-1}). $$

(48)

The mappings on the right-hand sides of (47) and (48) are the same. Moreover, by (27), the mapping between $(\bar{F}_i, \bar{F}_i +
where the last equality is by block independence. By the

\[ H(F_i | G_i) = \mathcal{H} + \delta_n \quad \text{and} \quad H(F_{i+1} | G_{i+1}) = \mathcal{H} - \delta_n. \]

Observe from (4d) and (4e) that Corollary 12 implies that there exists a nondecreasing sequence \( \delta_n \geq 0 \) such that

\[ H(F_i | G_i) = \begin{cases} \mathcal{H} + \delta_n, & i \in [\text{med}_{(n)}], \\ \mathcal{H} - \delta_n, & i \in [\text{med}_{(n)}], \end{cases} \] (49)

**Proof:** We show this using induction. The claim is true for \( n = 1 \) with \( \delta_0 = 0 \). For \( n > 1 \), we assume the claim is true for \( n - 1 \) and show it also holds for \( n \).

Let \( i \in [\text{med}_{(n)}] \) with base-vector \( \mathbf{b} \). Since \( n \geq 1 \), \( i \) is even (see Remark 1), and we denote \( i' = i/2 \). Let \( F_i, G_i \), as well as \( F_{i+1}, G_{i+1} \), be defined as in (43a) and let \( U_{i+1}, V_{i+1}, R_{i+1} \) be defined as in (44). We have, by (26) and (27),

\[ H(F_i | G_i) + H(F_{i+1} | G_{i+1}) = H(F_i, F_{i+1} | G_i, G_{i+1}) = H(U_{i+1}, V_{i+1} | G_i, G_{i+1}) = H(U_{i+1} | V_{i+1}, G_i, G_{i+1}). \] (50)

where the last equality is by block independence. By the induction assumption and stationarity there exists \( \delta_{n-1} \geq 0 \) such that

\[ H(U_{i'} | V_{i'}) = H(V_{i'} | R_{i'}) = \begin{cases} \mathcal{H} + \delta_{n-1}, & i' \in [\text{med}_{(n-1)}], \\ \mathcal{H} - \delta_{n-1}, & i' \in [\text{med}_{(n-1)}]. \end{cases} \]

Thus,

\[ H(F_i | G_i) + H(F_{i+1} | G_{i+1}) = 2\mathcal{H}. \] (51)

By (4d) and (4e) and since \( i \in [\text{med}_{(n)}] \), we have \( i + 1 \in [\text{med}_{(n)}] \). Recall from Remark 1 that since \( n \geq 1 \) then \( i \) is even and \( i + 1 \) is odd. By (26), (27), and since conditioning reduces entropy, we have

\[ H(F_{i+1} | G_{i+1}) \leq \min\{H(U_{i+1} | V_{i+1}), H(V_{i+1} | R_{i+1})\} \]

\[ = \mathcal{H} - \delta_{n-1}. \] (52)

From (51) and (52), we conclude that there must exist \( \delta_n \geq \delta_{n-1} \geq 0 \) such that \( H(F_i | G_i) = \mathcal{H} + \delta_n \) and \( H(F_{i+1} | G_{i+1}) = \mathcal{H} - \delta_n \). Finally, by Lemma 11, \( \delta_n \) must be independent of \( i \).

Recall that we wish to prove that the OT-BST is monopolarizing for the BI-process. From the proof of Corollary 12 it follows that \( \delta_n \geq \delta_{n-1} \) for any \( n \). This is not sufficient for monopolarization; to show monopolarization we must show that, unless we have already monopolarized, \( \delta_n > \delta_{n-1} + \Delta \) for some \( \Delta > 0 \) independent of \( n \). This is the role of Lemma 14 that follows. To this end, we will need an auxiliary lemma.

The binary entropy function \( h_2 \), defined in (1), is monotone increasing over \([0, 1/2]\). Denote the (cyclic) convolution of two numbers \( 0 \leq \alpha, \beta \leq 1/2 \) by

\[ \alpha * \beta = \alpha (1 - \beta) + \beta (1 - \alpha). \]

Since

\[ \alpha * \beta = \alpha + \beta (1 - 2 \alpha) = \beta + \alpha (1 - 2 \beta), \] (53)

we have \( h_2(\alpha * \beta) \geq h_2(\beta) \) for any \( \alpha, \beta \in [0, 1/2] \). More precisely, we have the following lemma; its proof can be found in Appendix C.

**Lemma 13.** Let \( 0 \leq \alpha, \beta \leq 1/2 \), \( a, b = 1, 2, \ldots, k \) and let \( p_a, q_b \geq 0 \) such that \( \sum_{a=1}^k p_a = \sum_{b=1}^k q_b = 1 \). If, for some \( \xi_1, \xi_2 > 0 \),

\[ \sum_{a=1}^k p_a h_2(a_\alpha) \geq \xi_1, \quad \sum_{b=1}^k q_b h_2(b_\beta) \leq \xi_2, \] (54)

then there exists \( \Delta(\xi_1, \xi_2) > 0 \) such that

\[ \sum_{a=1}^k \sum_{b=1}^k p_a q_b (h_2(a_\alpha * b_\beta) - h_2(b_\beta)) \geq \Delta(\xi_1, \xi_2). \]

Recall that \( i \in [\text{med}_{(n)}] \), with base-vector \( \mathbf{b} = [b_1, b_2] \), where \( b_1 \) and \( b_2 \) are of length \( 2^{r-1} \). Assume further that \( i \in [\text{med}_{(n)}] \), so that \( i \) is even, and \( i' = i/2 \). We define \( F_i, G_i \) as in (43a), and \( \tilde{U}_{i+1}, \tilde{V}_{i+1}, \tilde{R}_{i+1} \) as in (44).

**Lemma 14.** For all \( \xi > 0 \), if \( i \in [\text{med}_{(n)}] \) and

\[ H(U_{i+1} | \tilde{Q}_{i+1}), H(V_{i+1} | \tilde{R}_{i+1}) \in (\xi, 1 - \xi) \] (55)

then

\[ H(F_i | G_i) = \max\{H(U_{i+1} | \tilde{Q}_{i+1}), H(V_{i+1} | \tilde{R}_{i+1})\} \geq \Delta(\xi, 1 - \xi). \]

**Proof:** There is nothing to prove if \( \xi \geq 1/2 \). Therefore, we assume that \( \xi < 1/2 \). We show the proof for the case where \( H(V_{i+1} | \tilde{R}_{i+1}) \geq H(U_{i+1} | \tilde{Q}_{i+1}) \). The proof of the other case is similar and omitted.

We will use the simplified notation

\[ \tilde{p}(u, v, q, r) = \mathbb{P}(\tilde{U}_{i+1} = u, \tilde{V}_{i+1} = v, \tilde{Q}_{i+1} = q, \tilde{R}_{i+1} = r). \]

Since \( \tilde{U}_{i+1}, \tilde{Q}_{i+1} \) and \( \tilde{V}_{i+1}, \tilde{R}_{i+1} \) are independent, we have

\[ \tilde{p}(u, v, q, r) = \tilde{p}(u, q) \tilde{p}(v, r). \]

We also introduce the shorthand

\[ a_q = \min_u \tilde{p}(u, q), \quad b_r = \min_v \tilde{p}(v, r). \]

Recall that \( \tilde{U}_{i+1}, \tilde{V}_{i+1} \) are binary, so the minimizations are between two terms. As a result, \( 0 \leq a_q, b_r \leq 1/2 \). With this notation and by (55) we have

\[ H(U_{i+1} | \tilde{Q}_{i+1}) = \sum_q \tilde{p}(q) h_2(a_q) \geq \xi, \]

\[ H(V_{i+1} | \tilde{R}_{i+1}) = \sum_r \tilde{p}(r) h_2(b_r) \leq 1 - \xi. \]

Thus, by (45) and the independence of \( (\tilde{U}_{i+1}, \tilde{Q}_{i+1}) \) and \( (\tilde{V}_{i+1}, \tilde{R}_{i+1}) \), we obtain

\[ H(F_i | G_i) - H(V_{i+1} | \tilde{R}_{i+1}) = H(U_{i+1} + V_{i+1} | \tilde{Q}_{i+1}, \tilde{R}_{i+1}) - H(V_{i+1} | \tilde{R}_{i+1}) \]

\[ \geq \Delta(\xi, 1 - \xi). \]
We conclude that if ̃\( M \) of (38) further define ̃\( \xi \), then:

- if ̃\( \xi \leq 1/2 \) then \( H(̃\xi_i|G_i) < ξ, \forall i \in \text{med}_{\cdot}(n) \);
- if ̃\( \xi \geq 1/2 \) then \( H(̃\xi_i|G_i) > 1 - ξ, \forall i \in \text{med}_{\cdot}(n) \).

**Proof:** Define the indicator functions

\[
M_n^- = \begin{cases} 1, & H(̃\xi_i|G_i) > 1 - ξ, \forall i \in \text{med}_{\cdot}(n), \\ 0, & \text{otherwise}, \end{cases} \quad M_n^+ = \begin{cases} 1, & H(̃\xi_i|G_i) < ξ, \forall i \in \text{med}_{\cdot}(n), \\ 0, & \text{otherwise}, \end{cases} \quad M_n = \begin{cases} 1, & M_n^- = 1 \text{ or } M_n^+ = 1, \\ 0, & \text{otherwise}. \end{cases}
\]

Observe that \( M_n = 1 \) if and only if the OT-BST has \( (ξ, \text{med}_{\cdot}(n)), \text{med}_{\cdot}(n)) \)-monopolarized for the BI-process. Further define

\[ n_{th} = \min \{ n \in \mathbb{N} \mid M_n = 1 \}. \]

This is the first index \( n \) for which \( M_n = 1 \). We will show that \( n_{th} \) is finite by upper-bounding it.

By Corollary 12, there exists a nondecreasing sequence \( δ_n \geq 0 \) such that (49) holds. Since \( δ_n \) is a nondecreasing sequence, \( M_n = 1 \) for every \( n \geq n_{th} \). The entropy of a binary random variable is bounded between 0 and 1; thus for any \( n, 0 \leq ̃\xi - δ_n \leq ̃\xi + δ_n \leq 1 \). Hence, \( δ_n \leq \text{med}(̃\xi, 1 - ̃\xi) \).

We conclude that if ̃\( \xi \leq 1/2 \) and \( n \geq n_{th} \), then \( M_n^+ = 1 \), and if ̃\( \xi \leq 1/2 \) and \( n \geq n_{th} \), then \( M_n = 1 \). It now remains to upper-bound \( n_{th} \).

If \( M_0 = 1 \), then we may take \( n_{th} = 0 \) and we are done. Otherwise, we assume that \( M_0 = 0 \).

If, for some \( n \geq 0, M_n = 0 \), then by (49) and by definition of \( M_n^-, M_n^+ \), we obtain

\[ δ_n \leq ̃\xi - δ_n \leq ̃\xi + δ_n \leq 1 - ξ. \]

Rearranging, this yields

\[ M_n = 0 \Rightarrow δ_n \leq \min(̃\xi, 1 - ̃\xi) - ξ. \]

On the other hand, by (49) and Lemma 14, if \( M_{n-1} = 0 \) for some \( n \geq 1 \), we have

\[ ̃\xi + δ_n - (̃\xi + δ_{n-1}) = Δ(̃\xi, 1 - ̃\xi) \Rightarrow δ_n \geq δ_{n-1} + Δ(̃\xi, 1 - ̃\xi). \]

Continuing in this manner and recalling that \( δ_0 = 0 \), we obtain

\[ M_n = 0 \Rightarrow δ_n \geq nΔ(̃\xi, 1 - ̃\xi). \]

Now, let

\[ n_1 = 1 + \left\lfloor \frac{\min(̃\xi, 1 - ̃\xi) - ξ}{Δ(̃\xi, 1 - ̃\xi)} \right\rfloor, \]

and assume to the contrary that \( n_{th} > n_1 \). In particular, \( M_{n_1} = M_{n_1-1} = 0 \). Thus, by (56) and (57) we obtain

\[ n_1Δ(̃\xi, 1 - ̃\xi) \leq δ_{n_1} \leq \min(̃\xi, 1 - ̃\xi) - ξ. \]

Since \( Δ(̃\xi, 1 - ̃\xi) > 0 \), we rearrange and obtain

\[ n_1 \leq \frac{\min(̃\xi, 1 - ̃\xi) - ξ}{Δ(̃\xi, 1 - ̃\xi)}, \]

which contradicts (58) (see, e.g., [31, Equation 3.3]). We conclude that we must have \( n_{th} \leq n_1 \). We have found an upper bound for \( n_{th} \), thus completing the proof.

**Corollary 16.** Let \( L_0, M_0, \) and \( n_{th} \) be as in Proposition 15. Then, under the same setting as Proposition 15, for any \( 0 \leq ξ \leq 1 \) and \( n \geq n_{th} \), we have

- if \( ̃\xi \leq \frac{1-ξ}{2} \) then \( H(̃\xi_i|G_i) < ξ + ζ, \forall i \in \text{med}_{\cdot}(n) \).
- if \( ̃\xi \geq \frac{1-ξ}{2} \) then \( H(̃\xi_i|G_i) > 1 - ξ - ζ, \forall i \in \text{med}_{\cdot}(n) \).

**Proof:** This corollary follows from Proposition 15 and Corollary 12. Recall that by Corollary 12, there exists \( δ_n \geq 0 \) such that (49) holds.

We only prove the corollary for the case where \( ̃\xi \leq (1+ξ)/2 \). The case \( ̃\xi \geq (1-ξ)/2 \) is similar and omitted.

If \( ̃\xi \leq 1/2 \), we are done by Proposition 15. Otherwise, \( ̃\xi \geq 1/2 \), so by Proposition 15 and (49),

\[ i \in \text{med}_{\cdot}(n) \Rightarrow H(̃\xi_i|G_i) = ̃\xi + δ_n > 1 - ξ. \]

Rearranging, we obtain \( δ_n > 1 - ̃\xi - ξ \). Now, by (49),

\[ i \in \text{med}_{\cdot}(n) \Rightarrow H(̃\xi_i|G_i) = ̃\xi - δ_n < ̃\xi - (1 - ̃\xi - ξ) = ξ + 2̃\xi - 1 \leq ξ + (1 + ξ) - 1 = ξ + ξ, \]

where the final inequality is due to our assumption that \( ̃\xi \leq (1 + ξ)/2 \).

The upper bound for \( n_{th} \) given in Proposition 15 is pessimistic. It is based on the minimal change that must occur at every step of the OT-BST. The change at every OT-BST step is typically larger, and thus the actual required value of \( n_{th} \) is expected to be much smaller. We adapt [8, Proposition 2] to give better bounds on the required number of OT-BST steps to ensure monopolarization. To this end, we define, for \( y \in [0, 1] \) and \( x \in [0, \min\{y, 1 - y\}] \), the functions

\[ c(x, y) = h_2(h_2^{-1}(y + x) * h_2^{-1}(y - x) - y), \]
\[ d(x, y) = y - (y + x)(y - x), \]

where \( h_2^{-1} : [0, 1] \rightarrow [0, 1/2] \) is the inverse of \( h_2 \). Since \( h_2 \) is concave-in and increasing over \([0, 1/2] \), \( h_2^{-1} \) is convex-increasing and increasing over \([0, 1] \). We also define the sequence of functions

\[ C_0(y) = D_0(y) = 0, \]
\[ C_n(y) = c(C_{n-1}(y), y), \quad n = 1, 2, \ldots, \]
\[ D_n(y) = d(D_{n-1}(y), y), \quad n = 1, 2, \ldots. \]

**Lemma 17.** Let \( n \geq 0 \). If \( i \in \text{med}_{\cdot}(n) \) then

\[ C_n(̃\xi) \leq H(̃\xi_i|G_i) - ̃\xi \leq D_n(̃\xi). \]
If \( i \in [\text{med}_+(n)] \) then
\[
C_n(\tilde{H}) \leq \tilde{H} - H(\tilde{F}_i|G_i) \leq D_n(\tilde{H}).
\]

Proof: In light of Corollary 12, denote, for any \( n \geq 0 \) and arbitrary \( i \in [\text{med}_-(n)] \)
\[
\delta_n = H(\tilde{F}_i|G_i) - \tilde{H}.
\]
Observe that for arbitrary \( i \in [\text{med}_+(n)] \), by Corollary 12 we have \( \delta_n = \tilde{H} - H(\tilde{F}_i|G_i) \). Our goal is thus to show that for any \( n \geq 0 \),
\[
C_n(\tilde{H}) \leq \delta_n \leq D_n(\tilde{H}). \tag{59}
\]
The remainder of the proof mirrors the proof of [8, Proposition 2]. We prove the claim by induction. If \( n = 0 \), the claim is trivially true. Assume that the claim holds for some \( n \geq 0 \), and we will show it also holds for \( n + 1 \).

By block-independence of the BI-process we may use [6, Lemma 2.1], by which
\[
\tilde{H} + \delta_{n+1} \geq h_2^{-1}(\tilde{H} + \delta_n) + h_2^{-1}(\tilde{H} - \delta_n),
\]
\[
\tilde{H} + \delta_{n+1} \leq (\tilde{H} + \delta_n) + (\tilde{H} - \delta_n) - (\tilde{H} + \delta_n)(\tilde{H} - \delta_n).
\]
Rearranging, we obtain
\[
c(\delta_n, \tilde{H}) \leq \delta_{n+1} \leq d(\delta_n, \tilde{H}). \tag{60}
\]
Now, \( d(x, y) = x^2 - y^2 + y \) is increasing in \( x \) whenever \( x \geq 0 \). The function \( c(x, y) \) is also increasing for \( x \in [0, \min\{y, 1 - y\}] \). To see this, it suffices to show that \( c(x) = h_2^{-1}(x+y) + h_2^{-1}(x-y) \) is increasing, as \( h_2 \) is increasing. Denoting \( r(x) = h_2^{-1}(x) \) we obtain that
\[
\frac{dc}{dx}(x) = r'(y + x)(1 - 2r(y - x)) - r'(y - x)(1 - 2r(y + x)) \leq 0.
\]
where (a) is because \( r(\cdot) \) is increasing, and (b) is because \( r(\cdot) \) is convex so its derivative \( r'(\cdot) \) is increasing and since \( r(\cdot) \leq 1/2 \) by definition. Thus, by (60) and the induction hypothesis (59),
\[
\delta_{n+1} \geq c(\delta_n, \tilde{H}) \geq c(C_n(\tilde{H}), \tilde{H}) = C_{n+1}(\tilde{H}),
\]
\[
\delta_{n+1} \leq d(\delta_n, \tilde{H}) \leq d(D_n(\tilde{H}), \tilde{H}) = D_{n+1}(\tilde{H}),
\]
which completes the proof.

Example 5. Consider a BI-process with \( \tilde{H} = 0.2 \). We wish to find \( n_{th} \) that will ensure that the OT-BST is \((0.004, [\text{med}_+(n)], [\text{med}_-(n)])\)-monopolarizing for the BI-process whenever \( n \geq n_{th} \).

Proposition 15 gives the upper bound
\[
n_{th} \leq 1 + \left\lceil \frac{\tilde{H} - \xi}{\Delta(\xi, 1 - \xi)} \right\rceil = 40162.
\]
This is a prohibitive value. Thankfully, it is also unnecessarily pessimistic. To obtain a practical value for \( n_{th} \), we turn to Lemma 17, by which
\[
2.22 \cdot 10^{-5} \leq H(\tilde{F}_i|G_i) \leq 0.0041, \quad i \in [\text{med}_+(9)],
\]
\[
8.89 \cdot 10^{-6} \leq H(\tilde{F}_i|G_i) \leq 0.0031, \quad i \in [\text{med}_-(10)].
\]
Therefore, when \( \tilde{H} = 0.2, n_{th} = 10 \) suffices to ensure \((0.004, [\text{med}_+(n)], [\text{med}_-(n)])\)-monopolarization for \( n \geq n_{th} \).

C. Monopolarization for FAIM-derived Processes

We now show that the BST is monopolarizing for suitably chosen \( \eta, \zeta, \tilde{H} \) when applied to forgetful FAIM-derived s/o-processes. Our main goal is to establish Theorem 18 below.

Theorem 18. Let \( X_* \rightarrow Y_* \) be a forgetful FAIM-derived s/o-process. For every \( \eta > 0 \) there exist \( L_0, M_0 \), and \( n_0 \) such that if \( n \geq n_0 \) then a level-n BST initialized with parameters \( L_0 \) and \( M_0 \) is \((\eta, [\text{med}_+(n)], [\text{med}_-(n)])\)-monopolarizing.

Specifically, let \( F_{N_{n_0}} \rightarrow G_{N_{n_0}} \) be a transformed s/o-block of a level-n BST initialized with \( L_0 \) and \( M_0 \) as above. Then:

- if \( \zeta(X_*|Y_*) \leq 1/2 \) then \( H(F_i|G_i) < \eta, \quad \forall i \in [\text{med}_+(n)] \);
- if \( \zeta(X_*|Y_*) \geq 1/2 \) then \( H(F_i|G_i) > 1 - \eta, \quad \forall i \in [\text{med}_-(n)] \).

This theorem will follow as a corollary to Proposition 19 below. We will show in Proposition 19 that, when \( L_0 \) and \( M_0 \) are suitably chosen, there is a close relationship between the BST of a forgetful FAIM-derived s/o-process and the OT-BST of a BI-process. Since, by Proposition 15, the OT-BST of a BI-process is monopolarizing, this will imply that the BST is also monopolarizing.

The s/o-process \( X_* \rightarrow Y_* \) is a forgetful FAIM-derived s/o-process. By Lemma 6, it satisfies (33) with mixing sequences \( \psi_k, \phi_k \). We apply to s/o-block \( X_{n_0} \rightarrow Y_{n_0} \) a level-n BST initialized with parameters \( L_0 \) and \( M_0 \). The parameters \( L_0 \) and \( M_0 \) will be determined later. The BI-process \( X_* \rightarrow Y_* \) with parameter \( N_0 = 2L_0 + M_0 \) is defined as in Definition 11.

Recall our notation from Section IV-B for the BST and OT-BST. We will only consider medial indices. The BST is expressed using the sequence of functions \( f_{n,i}, g_{n,i} \), where \( i \in [\text{med}(n)] \). The OT-BST is expressed using the sequence of functions \( f_{n,i}, g_{n,i} \).

Let \( i \in [\text{med}(n)] \); its base-vector \( b \) is given by
\[
b = [b_1 \ b_2 \ \cdots \ b_{2^n}].
\]

We also denote
\[
a = [1 \ N_0 + 1 \ 2N_0 + 1 \ \cdots \ (2^n - 1)N_0 + 1],
\]
\[
z = [N_0 \ 2N_0 \ 3N_0 \ \cdots \ 2^n N_0].
\]

We further define for index \( i \in [\text{med}(n)] \):
\[
F_i = f_{n,i}(X_0), \quad G_i = g_{n,i}(X^b_{n,i} Y^a_{n,i}), \tag{61a}
\]
\[
\tilde{F}_i = \tilde{f}_{n,i}(X_0), \quad \tilde{G}_i = \tilde{g}_{n,i}(X^b_{n,i} Y^a_{n,i} L_0), \tag{61b}
\]
\[
\tilde{F}_i = \tilde{f}_{n,i}(\tilde{X}_0), \quad \tilde{G}_i = \tilde{g}_{n,i}(X^b_{n,i} Y^a_{n,i} L_0). \tag{61c}
\]

In words:
- \( F_i \Rightarrow G_i \) is a transformed s/o-pair obtained after applying a level-n BST to the FAIM-derived process;
- \( \tilde{F}_i \Rightarrow \tilde{G}_i \) is an OT-transformed s/o-pair obtained after applying a level-n OT-BST to the FAIM-derived process;
- \( \tilde{F}_i \Rightarrow \tilde{G}_i \) is an OT-transformed s/o-pair obtained after applying a level-n OT-BST to the BI-process.

Proposition 19. Fix \( n \geq 0, \epsilon_1 > 0, \text{ and } 0 < \epsilon_2 < \frac{1}{2} \). There exist \( L \) and \( M \) such that a level-n BST initialized with parameters \( L_0 \geq L, M_0 \geq M \) satisfies:
\[
|H(F_i|G_i) - H(F_i|\tilde{G}_i)| < 2\epsilon_1 + \sqrt{8\epsilon_2}. \tag{62}
\]
Proof: Denote
\[ \tilde{\Pi} = P_{X_{b.L_0}^b, y_{b.L_0}^b}, \]
\[ \hat{\Pi} = \prod_{\ell=1}^{2^n} P_{X_{b.L_0}^b, y_{b.L_0}^b}. \]

Then, \((X_{b.L_0}^b, y_{b.L_0}^b)\) is distributed according to \(\hat{\Pi}\) and \((X_{b.L_0}^b, y_{b.L_0}^b)\) is distributed according to \(\tilde{\Pi}\).

In Lemma 20 that follows we show that there exists \(L\) such that if \(L_0 \geq L\) then
\[ |H(\hat{F}_i) - H(\tilde{F}_i)| \leq 2\varepsilon_1. \]

Next, in Lemma 21 that follows we show that there exists \(M\) such that if \(M_0 \geq M\) then
\[ (1 - \varepsilon_2)\tilde{\Pi} \leq \hat{\Pi} \leq (1 + \varepsilon_2)\tilde{\Pi}. \]

This will enable us to use Lemma 22 below with \(f = \tilde{f}_{n,i}\) and \(g = \tilde{g}_{n,i}\) to obtain
\[ |H(\tilde{F}_i) - H(\hat{F}_i)| < \sqrt{8\varepsilon_2}. \]

Hence, we conclude that
\[ |H(\hat{F}_i) - H(\tilde{F}_i)| \leq |H(\tilde{F}_i) - F(\tilde{F}_i)| + |H(\hat{F}_i) - H(\tilde{F}_i)| \]
\[ \leq 2\varepsilon_1 + \sqrt{8\varepsilon_2}, \]

which completes the proof.

We now state and prove Lemmas 20 to 22.

**Lemma 20.** Fix \(n \geq 0\) and \(\varepsilon_1 > 0\). There exists \(L\) such that if \(L_0 \geq L\) then
\[ 0 \leq H(\hat{F}_i) - H(\tilde{F}_i) \leq 2\varepsilon_1. \]

Recall from Definition 10 that for a forgetful process, we may set the forgetfulness as small as desired by increasing the recollection. Moreover, for forgetful processes that satisfy Condition K, the forgetfulness decreases exponentially with the recollection (see Proposition 38 in Section VIII).

**Proof:** By (31), \(G_i \equiv (\tilde{G}_i, \hat{G}_i)\), where
\[ \tilde{G}_i = \tilde{g}_{n,i}(X_{b.L_0}^b, y_{b.L_0}^b), \]
\[ \hat{G}_i = (X_{b.L_0}^b, y_{b.L_0}^b). \]

Since \(\tilde{f}_{n,i} = \tilde{f}_{n,i}\), we have \(\hat{F}_i = \hat{F}_i\). Therefore,
\[ H(\hat{F}_i) = H(\hat{F}_i) = H(\tilde{F}_i) \]
where the inequality is because conditioning reduces entropy. This proves the left-hand side of (63).

We now turn to proving the right-hand side of (63). To this end, let \(\epsilon\) be the \(L\)-forgetfulness of the s/o-process; we soon specify how to set \(L\). Now, utilize Corollary 8 with \(\ell = b, \lambda = L\), and \(L_0 \geq L\) to obtain
\[ I(S_b; S^{b.L_0} | X_{b.L_0}^{b.L_0}, y_{b.L_0}^{b.L_0}) \leq 2^n \cdot 2\varepsilon. \]

We take \(L\) large enough so that \(\epsilon \leq \varepsilon_1 \cdot 2^{-n}\). Hence,
\[ 2\varepsilon_1 \geq I(S_b; S_{b.L_0} | X_{b.L_0}^{b.L_0}) \]
\[ \geq I(\tilde{F}_i; G_i; S_{b.L_0}, S_{b.L_0} | y_{b.L_0}^{b.L_0}) \]
\[ \geq I(\tilde{F}_i; S_{b.L_0} | S_{b.L_0}, G_i; y_{b.L_0}^{b.L_0}) \]
\[ \geq I(\tilde{F}_i; S_{b.L_0} | S_{b.L_0}, G_i) \]
\[ = H(\tilde{F}_i|G_i) - H(\tilde{F}_i|G_i, S_{b.L_0}, S_{b.L_0}) \]
\[ \geq H(\tilde{F}_i|G_i) - H(\tilde{F}_i|G_i, S_{b.L_0}, S_{b.L_0}) \]
\[ \geq H(\tilde{F}_i|G_i) - H(\tilde{F}_i|G_i, S_{b.L_0}, S_{b.L_0}) \]
\[ \geq H(\tilde{F}_i|G_i) - H(\tilde{F}_i|G_i). \]

Specifically, we have the Markov chain
\[ (S_{b.L_0}, S_{b.L_0}) \rightarrow (S_{b.L_0}, y_{b.L_0}^{b.L_0}) \rightarrow (F_i, G_i). \]

This completes the proof.

**Lemma 21.** Fix \(n \geq 0\) and \(\varepsilon_2 > 0\). There exists \(M\) such that if \(M_0 \geq M\) then
\[ P_{X_{b.L_0}^b, y_{b.L_0}^b} \leq (1 + \varepsilon_2) \prod_{\ell=1}^{2^n} P_{X_{b.L_0}^b, y_{b.L_0}^b}. \]

Proof: Recall that the mixing sequences of the original s/o-process \(X_\ast \rightarrow Y_\ast\) are \(\psi_\ast\) and \(\phi_\ast\), where \(\psi_\ast \geq 1\) is nonincreasing and \(\phi_\ast \leq 1\) is nondecreasing. By Lemma 6, both sequences approach 1 exponentially fast. Thus, we may choose \(M\) such that
\[ (\psi_\ast) \leq 1 + \varepsilon_2, \]
\[ (\phi_\ast) \geq 1 - \varepsilon_2. \]

For any \(M_0 \geq M\) we thus have
\[ (\psi_\ast^{M_0}) \leq 1 + \varepsilon_2, \]
\[ (\phi_\ast^{M_0}) \geq 1 - \varepsilon_2. \]
which is (65).

Remark over the same finite alphabet

Lemma 22. Let \( A \) and \( \tilde{A} \) be two discrete random variables over the same finite alphabet \( \mathcal{A} \). Denote \( \mathbb{P}(A = a) = p(a) \) and \( \mathbb{P}(\tilde{A} = a) = q(a) \) for all \( a \in \mathcal{A} \). Assume that for some \( 0 \leq \varepsilon < \frac{1}{6} \),

\[
(1 - \varepsilon)q(a) \leq p(a) \leq (1 + \varepsilon)q(a), \quad \forall a \in \mathcal{A}.
\] (68)

Then, for any \( f : \mathcal{A} \to \{0, 1\} \) and \( g : \mathcal{A} \to \mathcal{G} \), where \( \mathcal{G} \) is some finite alphabet, we have

\[
|H(f(A)|g(A)) - H(f(\tilde{A})|g(\tilde{A}))| < \sqrt{\varepsilon}. \]

We are now ready to prove Theorem 18.

Proof of Theorem 18: Choose \( \epsilon_1 > 0 \) and \( 0 < \epsilon_2 < \frac{1}{6} \) small enough such that

\[
\xi \triangleq \eta - 4\epsilon_1 - (2\epsilon_1 + \sqrt{8\epsilon_2}) > 0.
\] (69)

For example, one may take \( \epsilon_1 < \eta/12 \) and \( \epsilon_2 < \eta^2/32 \). Take \( n_{th} \) large enough so that Proposition 15 holds with \( \xi \) as above. Such \( n_{th} \) may be found using Lemma 17. Recall that Proposition 15 holds for any \( L_0 \) and \( M_0 \), so we are free to set them as desired.

By Proposition 19, for \( n_{th}, \epsilon_1, \epsilon_2 \) above, there exist \( L \) and \( M \) such that (62) holds for \( L_0 = L \) and \( M_0 = M \). That is,

\[
-(2\epsilon_1 + \sqrt{8\epsilon_2}) \leq H(F_i|G_i) - H(\tilde{F}_i|\tilde{G}_i) \leq (2\epsilon_1 + \sqrt{8\epsilon_2}).
\] (70)

In fact, we choose \( L_0 = L \) as in the proof of Lemma 20. This ensures that the \( L_0 \)-forlgetfulness of the s/o-process is upper-bounded by \( \epsilon_1 \). Thus, by Corollary 10, (41) holds with \( \varepsilon \leq \epsilon_1 \), so that

\[
-\epsilon_1 \leq \mathcal{H}(X_*|Y_*) - \tilde{H} \leq \epsilon_1.
\]

Hence, if \( 3\epsilon_1 \leq 1/2 \) then \( \tilde{H} \leq (1 + 4\epsilon_1)/2 \) and if \( \mathcal{H}(X_*|Y_*) \geq 1/2 \) then \( \tilde{H} \geq (1 - 4\epsilon_1)/2 \). Consequently, by Corollary 16 with \( \xi = 4\epsilon_1 \), if \( n \geq n_{th} \) then

\[
3\epsilon_1 \leq 1/2 \Rightarrow H(F_i|G_i) < \xi + 4\epsilon_1, \quad \forall i \in [\text{med}_i(n)],
\]

\[
\mathcal{H}(X_*|Y_*) \geq 1/2 \Rightarrow H(\tilde{F}_i|\tilde{G}_i) > 1 - \xi - 4\epsilon_1, \quad \forall i \in [\text{med}_i(n)].
\]

Combining the above with (69) and (70) we obtain that for \( n \geq n_{th} \),

\[
\mathcal{H}(X_*|Y_*) \leq 1/2 \Rightarrow H(F_i|G_i) < \xi, \quad \forall i \in [\text{med}_i(n)],
\]

\[
\mathcal{H}(X_*|Y_*) \geq 1/2 \Rightarrow H(\tilde{F}_i|\tilde{G}_i) > 1 - \eta, \quad \forall i \in [\text{med}_i(n)].
\]

This completes the proof.

VI. DECODING THE UNIVERSAL POLAR CODES

The universal polar codes consist of a concatenation of the BST and Arıkan’s polar codes. Ultimately, the codes consist of recursive applications of Arıkan transforms, which can be decoded efficiently using successive-cancellation decoding. The difference between the slow and fast stages lies in the order in which the Arıkan transforms are chained. Therefore, both the slow and fast polarization stages are decoded using successive-cancellation decoding, performed in lockstep.

Specifically, the decoder estimates the codeword bits in succession, assuming previous decoding decisions are correct. To decode a symbol, the decoder computes its likelihood ratio; this is performed recursively. If the symbol is “frozen,” the decoder returns its frozen value. In a non-symmetric case, this might employ some common randomness shared between the encoder and decoder, see [21] for details.
Due to the memory in the s/o-process, the recursive computation of likelihoods is done via the successive-cancellation trellis decoding of [15] and [16]. In this variation of successive-cancellation decoding, the decoder is cognizant of the existence of an underlying state connecting two blocks, and averages over it when computing likelihoods. This results in a slight increase in complexity; in a regular polar code, when there are $|S|$ states and the code length is $\hat{N}$, the decoding complexity is $O(|S|^3\hat{N}\cdot \log \hat{N})$, see [16, Theorem 2].

The overall codelength of the universal polar code is $N\cdot \hat{N}$ (see Section III-C), so its decoding complexity using successive-cancellation trellis decoding is $O(|S|^3N\hat{N}\cdot \log(N\hat{N}))$.

As mentioned in Section III-C, the overall decoding error of this scheme is upper-bounded by $N\hat{N} \cdot 2^{-\beta\hat{N}}$ for any $\beta < 1/2$ and $\hat{N}$ large enough.

VII. A CONTRACTION INEQUALITY

In this section we introduce a contraction inequality that will be useful in proving a sufficient condition for forgetfulness in Section VIII. To this end, we define a pseudo-metric $d$ between two nonnegative vectors that have the same support.

We will show that if a matrix $M$ satisfies a certain property called subrectangularity, then it has a parameter $\tau(M) < 1$ such that $d(x^TM,y^TM) \leq \tau(M)d(x,y)$.

This section invariably contains a large number of indices. For tractability, we adhere to the following notational conventions in this section. Indices $i$ and $k$ denote indices of rows of matrix $M$, and indices $j$, $l$ denote indices of columns of matrix $M$. Additionally, throughout this section, we implicitly assume that in any product of two matrices or a vector and a matrix, their dimensions match to enable forming these products.

Recall that the support $\sigma(x)$ of a vector $x$ is the set of its nonzero indices. That is, $\sigma(x) = \{i \mid x_i \neq 0\}$. The following pseudo-metric [34, Chapter 3.1], [35, Section 2] is defined for nonnegative vectors with the same support.

**Definition 12 (Projective distance).** Let $x$, $y$ be two nonnegative nonzero vectors such that $\sigma(x) = \sigma(y)$. The projective distance $d$ between the two vectors is $d(x,y) = \max \{j \mid x_j \neq 0, y_j \neq 0\} \cdot \max \{x_j/x_i, y_j/y_i \mid x_i \neq 0, y_i \neq 0\}$.

For row vectors we define $d(x^T,y^T) = d(x,y)$. If $x = y = 0$, we define $d(x,y) = 0$.

The projective distance is usually defined for positive vectors. Our definition generalizes it slightly for nonnegative vectors, provided they have the same support. In other words, we may assume that the (joint) zero indices of $x$ and $y$ are deleted before computing this distance. The projective distance is a pseudo-metric [34, Exercise 3.1]: it satisfies all of the properties of a metric over the nonnegative quadrant, with the exception that $d(x,y) = 0$ if and only if $x = cy$ for some $c > 0$.

The concept of a subrectangular matrix was introduced in [19] for square nonnegative matrices. However, it is easily extended to arbitrary nonnegative matrices. In this work, therefore, a subrectangular matrix need not be square. Subrectangularity will play a key role in the contraction inequality we develop.

**Definition 13 (Subrectangular matrix).** A nonnegative matrix $M$ is called subrectangular if $(M)_{i,j} \neq 0$ and $(M)_{k,l} \neq 0$ implies that $(M)_{i,l} \neq 0$ and $(M)_{k,j} \neq 0$.

We illustrate a subrectangular matrix in Figure 10. To better understand the meaning of this concept, in the following lemma we introduce equivalent characterizations of a subrectangular matrix. To this end, we remind the reader that a nonzero row (column) of a matrix contains at least one nonzero element, and that for a matrix $M$ we denote its set of nonzero rows by $N_r(M)$ and its set of nonzero columns by $N_c(M)$.

**Lemma 23.** Let $M$ be a nonnegative matrix. The following are equivalent:

1. The matrix $M$ is subrectangular.
2. If $M$ contains a zero element, either the entire row containing it or the entire column containing it are all zeros:

   \[ (M)_{i,j} = 0 \iff i \not\in N_r(M) \lor j \not\in N_c(M). \]  \hspace{1cm} (72)

3. The matrix $M$ satisfies

   \[ (M)_{i,j} \neq 0 \iff i \in N_r(M) \land j \in N_c(M). \]  \hspace{1cm} (73)

**Proof:** The second and third characterizations are clearly equivalent. Hence, it suffices to show that $1 \Rightarrow 2$ and $2 \Rightarrow 1$.

1 $\Rightarrow$ 2: Assume to the contrary that $M$ is subrectangular but (72) is not satisfied. That is, there exist $i,j$ such that $(M)_{i,j} = 0$ and $i \in N_r(M)$, $j \in N_c(M)$. Since row $i$ and column $j$ of $M$ are not all zeros, there exist $k,l$ such that $(M)_{i,k} \neq 0$ and $(M)_{k,j} \neq 0$. By subrectangularity of $M$, $(M)_{i,j}$ must also be nonzero, a contradiction.

2 $\Rightarrow$ 1: Assume that (73) holds. If $M$ is an all-zero matrix, or has just a single nonzero row (column), then $M$ is obviously subrectangular. Assume, therefore, that $M$ has at least two nonzero rows and at least two nonzero columns. That is, there exist $(i,j)$, $(k,l)$ such that $(M)_{i,j} \neq 0$ and $(M)_{k,l} \neq 0$. Thus, by (73), $i,k \in N_r(M)$ and $j,l \in N_c(M)$. Then, a second of use...
of (73) implies that \((M)_{i,l} \neq 0\) and \((M)_{k,j} \neq 0\). Therefore, \(M\) is subrectangular. □

Observe from (72) that if \(M\) is subrectangular and \(M'\) is obtained from \(M\) by multiplying some of its rows or columns by 0, then \(M'\) is also subrectangular. Similarly, if \(M''\) is obtained from \(M\) by deleting some of its rows or columns, then \(M''\) is also subrectangular. In particular, (73) implies that the matrix formed by deleting all of the all-zero rows and columns of \(M\) is positive — it contains only positive elements.

**Lemma 24.** If \(M\) is a nonzero subrectangular matrix and \(x, y\) are nonnegative vectors such that \(\|x^TM\|_1 > 0\) and \(\|y^TM\|_1 > 0\), then \(\sigma(x^TM) = \sigma(y^TM)\) and \(\sigma(Mx) = \sigma(My)\).

We remark that this lemma holds even if \(\sigma(x) \neq \sigma(y)\). In particular, it implies that if \(M\) is subrectangular and \(x, y\) are arbitrary nonnegative vectors such that \(x^TM\) and \(y^TM\) are nonzero, then \(d(x^TM, y^TM)\) is well-defined.

**Proof:** It suffices to prove the claim that \(\sigma(x^TM) = \sigma(y^TM)\), for the second claim follows by noting that \(M\) is subrectangular if and only if \(M^T\) is subrectangular. Without loss of generality, we may assume that \(M\) does not have all-zero rows. For, if it had such rows, we could remove them and delete the corresponding indices from \(x\) and \(y\) without affecting any of the values involved. This implies, by (72), that any column of \(M\) is either all positive or all zeros. Thus, for any nonzero row \(z\), we have \((z^TM)_i = 0\) if and only if column \(i\) of \(M\) is an all-zero column. The claim follows since both \(x\) and \(y\) are nonnegative and nonzero. □

The following corollary was stated as \([19, \text{Proposition 6.1}]\) without proof. We provide a short proof.

**Corollary 25.** If \(M\) is a subrectangular matrix and \(T, T'\) are some other nonnegative matrices (not necessarily subrectangular), then \(TM\) and \(T'M\) are subrectangular.

**Proof:** The case where either the matrix is trivial, so we assume they are both nonzero. It suffices to consider the case \(TM\), since that transpose of a subrectangular matrix remains subrectangular. By Lemma 24, every row of \(TM\) is either all-zeros, or has the same support as the other nonzero rows of \(TM\). This implies, by (73), that \(TM\) is subrectangular. □

We remark that a converse to Corollary 25 does not hold. That is, if a product of two nonnegative matrices is subrectangular, this does not imply that either of them is subrectangular. For example, if we denote by * an arbitrary positive value in a matrix, then \(T_1, T_2\) below are not subrectangular whereas their product \(T_1T_2\) is:

\[
T_1 = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix}, \quad T_2 = \begin{bmatrix} * & * \\ 0 & * \end{bmatrix}, \quad T_1T_2 = \begin{bmatrix} * & * \\ * & * \end{bmatrix}.
\]

We now introduce a parameter that plays a key role in the contraction inequalities we develop. To this end, recall that the support \(\sigma(M)\) of a matrix \(M\) is the set of index pairs

\[
\sigma(M) = \{(i, j) \mid i \in \mathcal{N}_r(M), j \in \mathcal{N}_c(M)\}.
\]

By (73), if \(M\) is subrectangular and \((i, j) \in \sigma(M)\) then \((M)_{i,j} > 0\).

**Definition 14 (Birkhoff contraction coefficient).** Let \(M\) be a nonnegative matrix. Its **Birkhoff contraction coefficient** \(\tau(M)\) is defined as follows.

- If \(M\) is subrectangular and nonzero, then
  \[
  \tau(M) \triangleq \sup_{x>0, y>0} \frac{d(x^TM, y^TM)}{d(x, y)}. \tag{74}
  \]
- If \(M\) is the zero matrix, then \(\tau(M) = 0\).
- If \(M\) is not subrectangular, then \(\tau(M) = 1\).

By Lemma 24 and the positivity of \(x\) and \(y\), the numerator of (74) is well-defined. That is, \(x^TM\) and \(y^TM\) have the same support. The denominator of (74) is also well-defined, as \(x\) and \(y\) are positive and thus have the same support as well. Finally, to ensure that the ratio in (74) is well-defined, we use the convention \(0/0 = 0\). Observe that the supremum in (74) is obtained for \(x \neq cy\) for \(c > 0\).

The Birkhoff contraction coefficient \([34, \text{Chapter 3}], [36]\) is usually defined for matrices with no all-zero columns. We generalize here the definition slightly to apply also to matrices with columns that are all-zeros. In light of Definition 12 and Lemma 24, the Birkhoff contraction coefficient of a matrix with some all-zero columns is simply the Birkhoff contraction coefficient of the matrix obtained by deleting its all-zero columns. We note in passing that

\[
\tau(M) = \tau(M^T), \tag{75}
\]

since \(d(x^TM, y^TM) = d(M^Tx, M^Ty)\).

The following theorem is a restatement of \([34, \text{Section 3.4}]\) (see \([36, \text{Theorem 1.1}]\) for an alternative proof).

**Theorem 26.** If \(M\) is subrectangular and nonzero, then

\[
\tau(M) = \frac{1 - \sqrt{\phi(M)}}{1 + \sqrt{\phi(M)}} < 1,
\]

where

\[
\phi(M) \triangleq \min_{i, j \in \mathcal{N}_r(M)} \frac{(M)_{i,j}(M)_{i,j}}{(M)_{i,j}^2} > 0. \tag{76}
\]

Since \(M\) is subrectangular and nonzero, all index pairs on the right-hand side of (76) are in the support of \(M\), by which \(\phi(M) > 0\). In other words, the Birkhoff contraction coefficient of a subrectangular matrix is the Birkhoff contraction coefficient of the positive matrix obtained by deleting all of its all-zero rows and columns. The proofs of this theorem in \([34, \text{Section 3.4}]\) and \([36, \text{Theorem 1.1}]\) assume no all-zero columns in \(M\). However, as explained after Definition 14, they hold without change for our slightly generalized definition of the Birkhoff contraction coefficient.

By Definition 14 and Theorem 26, if \(x\) and \(y\) are positive vectors and \(M\) is subrectangular, then

\[
d(x^TM, y^TM) \leq \tau(M)d(x, y).
\]

We now show that this holds in the more general case, where \(x\) and \(y\) are nonnegative vectors with the same support.

**Corollary 27.** If \(x, y\) are nonnegative vectors such that \(\sigma(x) = \sigma(y)\) and \(M\) is subrectangular, then

\[
d(x^TM, y^TM) \leq \tau(M)d(x, y). \tag{77}
\]
Proof: The claim is trivial if $x = y = 0$. If $x, y$ are positive, the claim follows from Definition 14 and Theorem 26. So, we assume that $x$ and $y$ are nonzero but have some zero elements. Denote by $\bar{x}, \bar{y}$ the vectors formed from $x, y$ by deleting their zero elements, and by $\bar{M}$ the matrix formed from $M$ by deleting the rows corresponding to these indices. The resulting vectors are positive and the resulting matrix remains subrectangular. Therefore,

$$d(x^T M, y^T M) = d(\bar{x}^T \bar{M}, \bar{y}^T \bar{M}) \leq \tau(\bar{M})d(\bar{x}, \bar{y}) = \tau(\bar{M})d(x, y).$$

Finally, observe that $(1 - \sqrt{x})/(1 + \sqrt{x})$ is a decreasing function of $x$ when $x \geq 0$; this is easily seen by computing its derivative, $-(\sqrt{x} - \sqrt{x^2 - 1})^{-1}$. Since $\bar{M}$ is formed from $M$ by deleting rows, $\phi(\bar{M}) \geq \phi(M)$. Thus, we must have $\tau(\bar{M}) \leq \tau(M)$, which completes the proof.

In the following lemma we prove an inequality, adapted from the proof of [35, Lemma 5], that is useful in the sequel.

**Lemma 28.** Let $\alpha_i > 0$, $\beta_i > 0$, and $\gamma_i \geq 0$ for all $i$. Assume that $\gamma_i > 0$ for some $i$. Then,

$$\min_{i} \frac{\alpha_i}{\beta_i} \leq \frac{\sum_i \gamma_i \alpha_i}{\sum_i \gamma_i \beta_i} \leq \max_{i} \frac{\alpha_i}{\beta_i}. \quad (78)$$

**Proof:** Denoting $\rho_i = \alpha_i/\beta_i$, we have

$$\frac{\sum_i \gamma_i \alpha_i}{\sum_i \gamma_i \beta_i} = \frac{\sum_i \gamma_i \rho_i}{\sum_i \gamma_i} = \sum_i \frac{\gamma_i \rho_i}{\sum_i \gamma_i} \rho_i = \sum_i \theta_i \rho_i,$$

where $\theta_i \geq 0$ for all $i$ and $\sum_i \theta_i = 1$. That is, the ratio on the left-hand side is a convex combination of the ratios $\rho_i$. Hence, it is lower-bounded by $\min_i \rho_i$ and upper-bounded by $\max_i \rho_i$, as required. $\blacksquare$

Armed with the above inequality, we can prove the following important property of the Birkhoff contraction coefficient.

**Lemma 29.** Let $M$ be a subrectangular matrix and let $T$ be a nonnegative matrix. Then,

$$\tau(TM) \leq \tau(M).$$

If, in addition, $T$ is subrectangular then

$$\tau(TM) \leq \tau(T)\tau(M). \quad (79)$$

**Remark 7.** Two remarks are in order. First, we note that (79) is adapted from [34, equation 3.7]. Second, there is nothing special about the ordering of the subrectangular and nonnegative matrix in the lemma. In particular, if the product $TM$ is replaced with the product $MT$ everywhere, the lemma holds unchanged. Indeed, by (75), $\tau(TM) = \tau((TM)^T) = \tau(M^T T^T)$ and $M$ is subrectangular if and only if $M^T$ is subrectangular.

**Proof:** There is nothing to prove if $TM = 0$, so we assume that $TM$ is nonzero.

By Corollary 25, $TM$ is subrectangular. For the first claim, let $i_0, l_0 \in N_i(TM)$ and $j_0, l_0 \in N_i(TM)$ achieve the minimum in (76); that is, be such that

$$\phi(TM) = \frac{(TM)_{i_0, j_0}}{(TM)_{i_0, l_0}} \leq \frac{(TM)_{j_0, l_0}}{(TM)_{j_0, j_0}}. \quad (73)$$

Thus, by (73), $(i_0, j_0), (k_0, l_0) \in \sigma(TM)$. This implies that $j_0, l_0 \in N_i(M)$ — otherwise, for example, we would have $(TM)_{i_0, l_0} = \sum_r (TM)_{r, l_0} = 0$, which contradicts $(i_0, j_0) \in \sigma(TM)$.

Hence,

$$\phi(TM) = \frac{\sum_i (TM)_{i_0, l_0} (TM)_{i_0, l_0}}{\sum_i (TM)_{j_0, l_0} (TM)_{j_0, j_0}} \leq \frac{\sum_i (TM)_{k_0, k_0}}{\sum_i (TM)_{k_0, j_0}} \leq \phi(M).$$

where (a) is by the left-hand inequality of (78), used twice and since $j_0, l_0 \in N_i(M)$ so that the subrectangularity of $M$; and in (b) we minimize over a set of indices that contains $j_0, l_0$. Having established $\phi(TM) \geq \phi(M)$ and, since $(1 - \sqrt{x})/(1 + \sqrt{x})$ is a decreasing function of $x$ for $x \geq 0$ (see the proof of Corollary 27), we conclude that $\tau(TM) \leq \tau(M)$.

For the second claim, if $T, M$ are both subrectangular, then for any positive $x, y$ we have $\sigma(x^T T) = \sigma(y^T T)$ and repeated applications of (77) yield

$$d(x^T TM, y^T TM) = d(x^T T M, y^T T M) \leq \tau(M)d(x^T T, y^T T) \leq \tau(M)\tau(T)d(x, y).$$

Thus, by (74), $\tau(TM) \leq \tau(T)\tau(M)$. $\blacksquare$

Applying Lemma 29 to a product of $m$ subrectangular matrices $M_1, M_2, \ldots, M_m$, we obtain

$$\tau(M_1 M_2 \cdots M_m) \leq \prod_{\ell=1}^m \tau(M_\ell). \quad (80)$$

Corollary 27 required that $x, y$ both have the same support. For the cases where $x$ and $y$ have different supports, we have the following lemma.

**Lemma 30.** Let $M$ be subrectangular and let $T$ be an arbitrary nonnegative matrix. Then, for any two nonnegative vectors $x$ and $y$ such that $\left\|x^T TM\right\|_1 > 0$ and $\left\|y^T TM\right\|_1 > 0$,

$$d(x^T TM, y^T TM) \leq 4\ln \left(\frac{1 + \tau(M)}{1 - \tau(M)}\right). \quad (81)$$

Since $M$ is subrectangular, $\tau(M) < 1$, which implies that the right-hand side of (81) is finite.

**Proof:** There is nothing to prove if $TM = 0$, so we assume that $TM$ is nonzero. By Corollary 25, $TM$ is subrectangular.

Fix any $i_0 \in N_i(M)$. Such an $i_0$ must exist because $M$ is subrectangular and $x^T M$ is nonzero by assumption. By Lemma 24, and subrectangularity of $M$,

$$\sigma(e_{i_0}^T M) = \sigma(x^T M) = \sigma_{\text{NW}}(\bar{M}). \quad (82)$$

\]
where \( \phi \) is defined in (76). The right-hand equality follows directly from Theorem 26, so we concentrate on proving the inequality.

For any \( j \in N_c(\tilde{M}) \) denote
\[
\rho_j = \frac{(e_0^T \tilde{M})_j}{(x^T \tilde{M})_j} = \sum_{k \in N_c(\tilde{M})} \frac{x_k(\tilde{M})_{k,j}}{(\tilde{M})_{k,j}}.
\]

The denominator is positive by (82), so \( \rho_j \) is well-defined. Now, for \( j, l \in N_c(\tilde{M}) \)
\[
\frac{\rho_j}{\rho_l} = \max_{k \in N_c(\tilde{M})} \frac{(\tilde{M})_{k,l}}{(\tilde{M})_{k,j}},
\]
where (a) is by Lemma 28 and in (b) we maximize over a set that contains \( i_0 \).

Hence, recalling the definition of the projective distance, (71),
\[
d(e_0^T \tilde{M}, x^T \tilde{M}) = \ln \max_{j, l \in N_c(\tilde{M})} \frac{\rho_j}{\rho_l}
\]
\[
\leq \max_{i, j, l \in N_c(\tilde{M})} \frac{(\tilde{M})_{i,j}}{(\tilde{M})_{i,l}} \cdot \frac{(\tilde{M})_{k,l}}{(\tilde{M})_{k,j}}
\]
\[
= \frac{1}{\phi(\tilde{M})}.
\]

By the symmetry and triangle inequality properties of the projective distance [34, Exercise 3.1],
\[
d(x^T \tilde{M}, y^T \tilde{M}) \leq d(e_0^T \tilde{M}, x^T \tilde{M}) + d(e_0^T \tilde{M}, y^T \tilde{M}).
\]

Thus, by Lemma 29 and since \( \ln((1 + x)/(1 - x)) \) is monotone increasing for \( 0 \leq x < 1 \), (81) will follow if we show that
\[
d(e_0^T \tilde{M}, x^T \tilde{M}) \leq \ln \left( \frac{1}{\phi(\tilde{M})} \right) = 2 \ln \left( \frac{1 + \tau(\tilde{M})}{1 - \tau(\tilde{M})} \right),
\]
where \( \phi \) is defined in (76). The right-hand equality follows directly from Theorem 26, so we concentrate on proving the inequality.

For any \( j \in N_c(\tilde{M}) \) denote
\[
\rho_j = \frac{(e_0^T \tilde{M})_j}{(x^T \tilde{M})_j} = \sum_{k \in N_c(\tilde{M})} \frac{x_k(\tilde{M})_{k,j}}{(\tilde{M})_{k,j}}.
\]

The denominator is positive by (82), so \( \rho_j \) is well-defined. Now, for \( j, l \in N_c(\tilde{M}) \)
\[
\frac{\rho_j}{\rho_l} = \max_{k \in N_c(\tilde{M})} \frac{(\tilde{M})_{k,l}}{(\tilde{M})_{k,j}},
\]
where (a) is by Lemma 28 and in (b) we maximize over a set that contains \( i_0 \).

Hence, recalling the definition of the projective distance, (71),
\[
d(e_0^T \tilde{M}, x^T \tilde{M}) = \ln \max_{j, l \in N_c(\tilde{M})} \frac{\rho_j}{\rho_l}
\]
\[
\leq \max_{i, j, l \in N_c(\tilde{M})} \frac{(\tilde{M})_{i,j}}{(\tilde{M})_{i,l}} \cdot \frac{(\tilde{M})_{k,l}}{(\tilde{M})_{k,j}}
\]
\[
= \frac{1}{\phi(\tilde{M})}.
\]

**Proof:** Since \( \|x^T M_i^m\|_1 > 0 \), we conclude that \( x^T M_i^m \) is nonzero for any \( 1 \leq s \leq m \), and the same holds if we replace \( x \) with \( y \). Thus, the left-hand side of (84) is well-defined. We will show that
\[
\ln \left( \frac{\|x^T M_i^m\|_1}{\|x^T M_i^m\|_1} \right) = \ln \left( \frac{\|y^T M_i^m\|_1}{\|y^T M_i^m\|_1} \right) \leq d(\tilde{x}, \tilde{y}) \cdot \prod_{\ell=2}^m \tau(M_\ell).
\]

The right-hand side is well-defined since, by Corollary 25, \( M_r \) is subrectangular for any \( 1 \leq r \leq s \leq m \) and by Lemma 24. Then, as \( \sigma(x) = \sigma(y) \) by Lemma 24, (84) will follow from Corollary 27 and (80).

Denote \( J = \sigma(\tilde{x}^T M_i^m) = \sigma(\tilde{y}^T M_i^m) = N_c(\tilde{M}_m) \), where the equalities are by Lemma 24 and subrectangularity. By the right-hand inequality of (78),
\[
\|y^T M_i^m\|_1 = \|\tilde{y}^T M_i^m\|_1 = \|\tilde{y}^T M_i^m\|_1 \leq \max_{i \in J} (\tilde{y}^T M_i^m)_{ij}.
\]

Next, denote by \( t_j = \|(T)_{ij}\| \) the sum of the \( j \)th row of \( T \). Since \( T \) is nonzero, \( t_j > 0 \) for some \( j \). Thus, a second application of the right-hand inequality of (78) yields
\[
\|y^T M_i^m\|_1 = \|\tilde{y}^T M_i^m\|_1 \leq \max_{j \in J} (\tilde{y}^T M_i^m)_{ij} \leq \max_{j \in J} (\tilde{y}^T M_i^m)_{ij}.
\]

Combining, we obtain
\[
\|x^T M_i^m\|_1 = \|\tilde{x}^T M_i^m\|_1 = \|\tilde{x}^T M_i^m\|_1 \leq \max_{j \in J} (\tilde{x}^T M_i^m)_{ij} = \|(T)_{ij}\|.
\]

Taking the logarithm of both sides, the right-hand side becomes \( d(\tilde{x}^T M_i^m, \tilde{y}^T M_i^m) \), which completes the proof.

Combining the above results we obtain the following corollary.

**Corollary 32.** Let \( M_1, M_2, \ldots, M_m \) be a sequence of square nonzero subrectangular matrices, and let \( T \), as well as \( T_1, T_2, \ldots, T_m \) be arbitrary square nonnegative and nonzero matrices. Denote
\[
R = T_1M_1T_2 \cdots T_mM_m.
\]

Then, for any two nonnegative nonzero vectors \( x, y \) such that \( \|x^T RT\|_1 > 0 \) and \( \|y^T RT\|_1 > 0 \) we have
\[
\ln \left( \|x^T RT\|_1 \right) = \ln \left( \|y^T RT\|_1 \right) \leq \ln \left( \frac{\|x^T RT\|_1}{\|y^T RT\|_1} \right) \leq 4 \ln \left( \frac{1 + \tau(M_1)}{1 - \tau(M_1)} \right) \cdot \prod_{\ell=2}^m \tau(M_\ell).
\]

**Proof:** The claim follows from Corollary 25, Lemmas 29 and 30, and Proposition 31.

**Observe** that (85) remains true if we replace ‘\( \ln \)’ with ‘\( \log \)’.

**Discussion.** Our Proposition 31 and Corollary 31 and Corollary 32 generalize [18, Theorem 2] in several ways. In [18, Theorem 2], the matrices \( M_1, \ldots, M_m, T \) are all strictly positive. Each matrix corresponds to an observation of a hidden Markov model \( (A_n, B_n) \), where the \((i, j)\) item of the matrix that corresponds to observation \( b \in \mathcal{B} \) is the probability that \( A_{n+1} = j \) and \( B_{n+1} = b \) given that \( A_n = i \). In particular, [18, Theorem 2] assumes that every observation \( b \in \mathcal{B} \) can be emitted from the same number of
states \(a \in \mathcal{A}\),\(^7\) and that it is possible to transition between any two states of \(\mathcal{A}\) in one step. In this work, however, we are not confined to such assumptions. Our formulation allows for each observation to originate from a different number of states. Moreover, our formulation does not assume that one can move from every state of \(\mathcal{A}\) to every other state of \(\mathcal{A}\) in one step.

VIII. HIDDEN MARKOV MODELS THAT FORGET THEIR INITIAL STATE

In this section we show that hidden Markov models that satisfy a mild requirement forget their initial state. Specifically, we will consider the mutual information between the state at time \(n+1\) and the model’s initial state given the observations in between. The contraction inequality of Section VII will enable us to show that this mutual information vanishes with \(n\). This enables us to obtain a sufficient condition — Condition \(K\) — for forgetfulness. The development in this section is based on the techniques of [19].

A. Hidden Markov Models

A hidden Markov model (HMM) is a process \((A_n, B_n)\), where \(A_n \in \mathcal{A}\) is a Markov chain and \(B_n \in \mathcal{B}\) is an observation that is a function of \(A_n\). The alphabets \(\mathcal{A}\) and \(\mathcal{B}\) are assumed finite. Without loss of generality, \(\mathcal{A} = \{1, 2, \ldots, |\mathcal{A}|\}\) and \(\mathcal{B} = \{1, 2, \ldots, |\mathcal{B}|\}\). A detailed description of the setting we consider follows.

Let \(A_n, n \in \mathbb{Z}\) be a homogeneous Markov process assuming values in some finite alphabet \(\mathcal{A}\). Denote by \(p(j|i)\) its transition probability function, which is independent of the time index \(n\). That is,

\[
p(j|i) = \mathbb{P}(A_n = j | A_{n-1} = i), \quad i, j \in \mathcal{A}.
\]

The \(|\mathcal{A}| \times |\mathcal{A}|\) transition probability matrix \(M\) of the Markov chain is defined by

\[
(M)_{i,j} = p(j|i), \quad i, j \in \mathcal{A}.
\]

This is a stochastic matrix: \((M)_{i,j} \geq 0\) for all \(i, j \in \mathcal{A}\) and for any \(i\), \(\sum_j (M)_{i,j} = 1\). We assume that the process \(A_n\) is aperiodic and irreducible (in some literature such Markov chains are called regular). That is, we assume that the matrix \(M\) is aperiodic and irreducible (see, e.g., [32, Proposition 4.1]). This implies [32, Theorems 1.9 and 4.2] that the process has a unique stationary distribution \(\pi\), which is positive.

Let \(f : \mathcal{A} \to \mathcal{B}\) be a deterministic function. For simplicity, we assume that \(\mathcal{B}\) is finite. An observation of \(A_n\) is \(B_n = f(A_n)\). Denote, for any set \(B \subseteq \mathcal{B}\),

\[
f^{-1}(B) = \{i \in \mathcal{A} | f(i) = b, b \in B\}.
\]

Then, \(\mathbb{P}(B_n = b) = \mathbb{P}(A_n \in f^{-1}(b))\). We assume that \(\mathcal{B}\) contains only observations that actually appear, that is, \(\mathcal{B} = \{b | f(i) = b, i \in \mathcal{A}\}\).

The process \((A_n, B_n)\) described above is called a hidden Markov model. We summarize this in the following definition.

\begin{definition}[Hidden Markov model] Let \(A_n\) be a homogeneous Markov process taking values in \(\mathcal{A}\) with transition probability matrix \(M\), which is aperiodic and irreducible. Let \(f : \mathcal{A} \to \mathcal{B}\) be a deterministic function, and let \(B_n = f(A_n)\). The process \((A_n, B_n)\) is called a hidden Markov model. Additionally, we use the following terminology:

- \(A_n\) is the state of the process,
- \(B_n\) is the observation of the process.
\end{definition}

Typically, multiple states would have the same observation. That is, for \(b \in \mathcal{B}\), the set \(f^{-1}(b)\) typically contains multiple elements. The actual state of the process is hidden, and the observation provides only partial information on the state.

The restriction to a deterministic function \(f\), rather than a probabilistic one, seemingly presents a limitation. However, in appendix E we show that there is no loss of generality in assuming that \(f\) is deterministic. That is, we show that the deterministic and probabilistic settings are equivalent. We emphasize that taking the viewpoint of deterministic \(f\) is done for convenience and to facilitate the derivation that follows. In particular, in our setting of a FAIM process, \((S_n, X_n, Y_n)\), without loss of generality one may assume that \((X_n, Y_n)\) is a deterministic function of the state \(S_n\).

The following notation, taken from [19], will be useful. Define the matrices \(M(b), b \in \mathcal{B}\), by

\[
(M(b))_{i,j} = \begin{cases} p(j|i), & \text{if } f(j) = b \\ 0, & \text{otherwise} \end{cases}
\]

In words, \((M(b))_{i,j}\) is the probability of transitioning from state \(i \in \mathcal{A}\) to state \(f(i) \in \mathcal{A}\) and observing \(b \in \mathcal{B}\) after having arrived at state \(j\). That is,

\[
(M(b))_{i,j} = \mathbb{P}(A_n = j, B_n = b | A_{n-1} = i).
\]

For a sequence of observations \(b^n_r, r \leq s\), we denote

\[
M(b^n_r) \triangleq M(b_{r+1}) \cdots M(b_s).
\]

We call \(\tau(M(b^n_r))\) the Birkhoff contraction coefficient induced by the sequence \(b^n_r\).

The matrices \(M(b)\) are nonzero and substochastic — they are nonnegative with unequal row sums, all less than or equal to 1. We can reconstruct \(M\) from them using

\[
M = \sum_b M(b).
\]

Example 8 in appendix E shows the matrix \(M\) and its decomposition to matrices \(M(b)\) for a specific channel with memory.

We also define for any \(a \in \mathcal{A}\) the matrix \(\mathbb{I}_a\) by

\[
(\mathbb{I}_a)_{i,j} = \begin{cases} 1, & \text{if } i = j = a \\ 0, & \text{otherwise} \end{cases}
\]

This matrix has a single nonzero element: ‘1’ on the diagonal, at the \((a, a)\) position.

The process \((A_n, B_n)\) is completely characterized by the matrices \(M(b), b \in \mathcal{B}\), and its initial distribution. We assume

\footnote{We note that the authors of [18] claim that this assumption can be relaxed with an appropriate extension, but they omit it and its derivation.}
that the process is stationary, so its initial distribution is \( \pi \), its unique stationary distribution. Thus, \( (\pi)_i = P(A_0 = i) \) and

\[
P(B_1 = b_1) = \sum_{j \in A} P(A_1 = j, B_1 = b_1)
= \sum_{i,j \in A} P(A_1 = j, B_1 = b_1 | A_0 = i) P(A_0 = i)
= \|\pi^T M(b_1)\|_1.
\]

Moreover, the probability of observing the sequence \( b^n_1 \) is given by [19, Lemma 2.1]

\[
P(B^n_1 = b^n_1) = \|\pi^T M(b^n_1)\|_1
= \|\pi^T M(b_1)^n M(b_2) \cdots M(b_n)\|_1.
\] (87)

Similarly, for any \( a \in A \),

\[
P(A_n = a, B^n_1 = b^n_1) = (\pi^T M(b^n_1))_a = \|\pi^T M(b^n_1)\|_1
\]

and

\[
P(A_{n+1} = a, B^n_1 = b^n_1) = (\pi^T M(b^n_1)M)_a
= \|\pi^T M(b^n_1)M\|_1
= \|\pi^T M(b^n_1)T_a\|_1.,
\] (88)

where we denoted for any \( a \in A \),

\[T_a \triangleq M_{i,a}.
\]

When \( P(B^n_1 = b^n_1) > 0 \) we further have by (87) and (88),

\[
P(A_{n+1} = a | B^n_1 = b^n_1) = \frac{P(A_{n+1} = a, B^n_1 = b^n_1)}{P(B^n_1 = b^n_1)}
= \|\pi^T M(b^n_1)T_a\|_1.
\] (89)

This is well-defined because if \( P(B^n_1 = b^n_1) > 0 \) then the denominator on the right-hand side of (89) must also be positive.

Let us now consider the case where the initial state of the process is known. In this case, \( P(B_1 = b_1 | A_0 = a_0) = \|e_{a_0}^T M(b_1)\|_1 \). Similar to the above, we obtain

\[
P(B^n_1 = b^n_1 | A_0 = a_0) = \|e_{a_0}^T M(b^n_1)\|_1.
\] (90)

\[
P(A_{n+1} = a | B^n_1 = b^n_1, A_0 = a_0) = \|e_{a_0}^T M(b^n_1)T_a\|_1
= \|e_{a_0}^T M(b^n_1)\|_1.
\] (91)

provided that the probability in (90) is positive.

In (87)–(91), we have computed probabilities for particular realizations of \( A_0, B^n_1 \) and \( A_{n+1} \). Generally, however, these are random variables. They are jointly generated as follows. First, draw \( A_0 \) according to \( \pi \). Then, at time \( n \), draw \( A_n \) according to the \( A_{n-1} \)th row of \( M \) and compute \( B_n = f(A_n) \).

These random variables give rise to the random variables \( P(A_{n+1} | B^n_1) \) and \( P(A_{n+1} B^n_1 A_0) \), obtained by plugging \( A_{n+1}, B^n_1 \), and \( A_0 \) for \( a, b^n_1 \), and \( a_0 \) respectively in the right-hand sides of (89) and (91). They are well-defined with probability 1. In other words, we can always compute their values via (89) and (91); with probability 0 will the denominators on the right-hand sides of these equations equal 0. These random variables are of interest because

\[
I(A_0; A_{n+1} | B^n_1) = \mathbb{E} \left[ \log \frac{P(A_{n+1} B^n_1 | A_0)}{P(A_{n+1} | B^n_1)} \right].
\] (92)

Using (89) and (91) we write this as

\[
I(A_0; A_{n+1} | B^n_1)
= \mathbb{E} \left[ \log \left( \frac{\|e_{a_0}^T M(b^n_1)\|_1}{\|\pi^T M(b^n_1)\|_1} \right) \right].
\] (93)

As above, the argument of the expectation is well-defined with probability 1.

The Markov chain \( A_n \) is finite-state, irreducible, and aperiodic. A classic result on such Markov chains [32, Theorem 4.3], [37, Theorem 8.9], which harks back to the days of A. A. Markov [38], is that the chain approaches its stationary distribution exponentially fast, regardless of its initial state. In particular, this implies that \( I(A_0; A_{n+1}) \rightarrow 0 \) as \( n \rightarrow \infty \). By the Markov property we also have \( I(A_0; A_{n+1} A^n_1) = 0 \). Our setting, however, is a hidden Markov setting, and we will be interested in whether \( I(A_0; A_{n+1} | B^n_1) \rightarrow 0 \). In general, the answer to this is negative — even when \( A_n \) is finite-state, aperiodic, and irreducible — see Example 3 in Section V-A, above.8

Our goal in the next subsection is to show that under a certain Condition K, \( I(A_0; A_{n+1} | B^n_1) \rightarrow 0 \) as \( n \rightarrow \infty \). This will employ (85), which bounds expressions of a form similar to the argument of the expectation in (93).

Remark 8. An expression similar to (92) was pointed out in [18, Equation 3.7], in the proof of [18, Theorem 2]. There, the goal was to show that \( I(A_0; B_n | B^n_{1-1}) \rightarrow 0 \). This was done under a restrictive assumption that transitions between any two states in one step may happen with strictly positive probability. When put in our notation, this implies that the matrices \( M(b), b \in \mathbb{B} \), contain only two types of columns: strictly positive columns and zero columns.9 In this case, the matrices \( M(b) \) are all subrectangular, so their Birkhoff contraction coefficients are strictly less than 1; this allows one to use (85) directly (with \( T_e = \mathbb{I} \) for all \( e \)) and obtain that the mutual information indeed vanishes as \( n \) grows. In this paper, we alleviate this restrictive assumption, and allow for a more general scenario where the individual matrices \( M(b) \) may also be not subrectangular.

We further remark that, by the data processing inequality (2), \( I(A_0; A_{n+1} | B^n_1) \rightarrow 0 \) implies that \( I(A_0; B_{n+1} | B^n_1) \rightarrow 0 \).

B. Forgetting the Initial State

We now show that under the following Condition K (so named in honor of Prof. Thomas Kailjser who had first suggested it in [19]), the mutual information \( I(A_0; A_{n+1} | B^n_1) \) vanishes with \( n \).

8Where the state is \( A_n = S_n \) and the observation is \( B_n = Y_n \). It can be shown [19, Section 10] that this HMM does not satisfy Condition K.

9The assumption of [18] is that \( B \) is positive. Since \( M(b) \) is comprised of columns of \( M \) and zero columns, any nonzero column of \( M(b) \) must be positive.
Condition K. The HMM \((A_n, B_n)\) is characterized by matrices \(M(b), b \in \mathcal{B}\) such that:

1. The matrix \(M = \sum_{b \in \mathcal{B}} M(b)\) is aperiodic and irreducible.
2. There exists an ordered sequence \(\beta_1, \beta_2, \ldots, \beta_k\) of elements of \(\mathcal{B}\) such that the matrix \(M(\beta_1)M(\beta_2) \cdots M(\beta_k)\) is non-zero and subrectangular.

The following are all examples where it is easy to check by inspection that Condition K is satisfied:

- the transition matrix \(M\) is positive (or, more generally, subrectangular);
- there exists an observation \(\beta\) for which \(M(\beta)\) has just a single column;
- there exists an observation \(\beta\) for which \(M(\beta)\) is subrectangular.

Generally, though, inspection may not suffice to declare that Condition K is satisfied.

Remark 9. The ability of a hidden Markov model to “forget” its initial state has also been studied under somewhat weaker assumptions than Condition K. The interested reader is invited to consult [39, 40]. It may be possible to generalize the results of this paper to processes that satisfy these weaker assumptions and do not satisfy Condition K. We leave such endeavors to future work.

Theorem 33. Suppose the HMM \((A_n, B_n)\) satisfies Condition K. Then, for every \(\epsilon > 0\) there exists an integer \(\lambda\) such that if \(n \geq \lambda\) then

\[
I(A_0; A_{n+1}|B^n_n) \leq \epsilon.
\]

The proof is given in the next subsection, and will follow from Proposition 38, which provides a characterization of the rate at which the mutual information vanishes. The idea is to use techniques similar to the ones used in the study of recurrence times of Markov chains. Namely, we bound the probability that in a long sequence of observations there will be sufficient non-overlapping occurrences of sequences that induce a Birkhoff contraction coefficient below a certain threshold. Armed with this bound, we employ Corollary 32 in (93) to obtain an upper bound on the mutual information.

Example 6. Let \(A_n\) be a finite-state Markov chain with irreducible and aperiodic transition probability matrix \(M\). Consider the case of no observations: \(B_n = 0\) regardless of \(A_n\). In this case, \(M(0) = M\) and Condition K is satisfied, as there exists \(k_0\) such that \(M^{k_0} > 0\) [34, Theorem 1.4]. Therefore, by Theorem 33, we have \(I(A_0; A_{n+1}) \to 0\) as \(n \to \infty\). As mentioned above, this is a well-known result for finite-state, irreducible, and aperiodic Markov chains. We note in passing that there exist other information-theoretic proofs that \(I(A_0; A_{n+1}) \to 0\) as \(n \to \infty\), see, e.g., [41].

Corollary 34. Suppose the HMM \((A_n, B_n)\) satisfies Condition K. Then, for every \(\epsilon > 0\) there exists an integer \(\lambda\) such that if \(n \geq \lambda\) then

\[
I(A_1; A_n|B^n_n) \leq \epsilon.
\]

Proof: The conditions of Theorem 33 are satisfied. Let \(\lambda\) be such that \(I(A_1; A_n|B^n_n) \leq \epsilon\) for any \(n \geq \lambda\).

We first show (94). Recall that \(B_j\) is a function of \(A_j\) for any \(j\). Thus, for any \(n \geq \lambda\), \(I(A_1, B_1; A_n, B_n|B^n_n) = I(A_1; A_n|B^n_n) \leq \epsilon\). Therefore,

\[
\epsilon \geq I(A_1, B_1; A_n, B_n|B^n_n) = I(B_1; A_n, B_n|B^n_n) + I(A_1; B_n|B^n_n) + I(A_1; A_n|B^n_n).
\]

Thus, since mutual information is nonnegative, each of the summands on the right-hand side is upper-bounded by \(\epsilon\). This yields (94).

To see (95), since \(A_0 \not\rightarrow (A_1, B_1) \not\rightarrow A_n\), we use (2) and obtain

\[
I(A_0; A_n|B^n_n) \leq I(A_1; A_n|B^n_n) \leq \epsilon,
\]

as required.}

C. Proof of Theorem 33

The goal of this subsection is to prove Theorem 33. To this end, we make the following definition.

Definition 16. Let \(n^*\) be a positive integer, and \(\delta^*, \tau^* \in (0, 1)\). The HMM \((A_n, B_n)\) is called an \((n^*, \delta^*, \tau^*)\)-HMM if it satisfies

\[
\mathbb{P} (\tau(M(B^n_n)) \leq \tau^*|A_0 = a_0) \geq 1 - \delta^*, \quad \forall a_0 \in A.
\]

In words, the HMM has a probability at least \((1 - \delta^*)\) of emitting by time \(n^*\) an observation sequence that induces a Birkhoff contraction coefficient at most \(\tau^*\), regardless of its initial state.

We say that an HMM is a KHMM if it is an \((n^*, \delta^*, \tau^*)\)-HMM for some \((n^*, \delta^*, \tau^*)\).

\[10\text{Taking } A_n = S_n, B_n = (X_n, Y_n), \text{ and } C_n = Y_n. \]
Observe that if \((A_n, B_n)\) is an \((n_*, \delta_*, \tau_*)\)-KHM and \(n_* \leq n'_*, \delta_* \leq \delta'_*, \text{ and } \tau_* \leq \tau'_*\) then \((A_n, B_n)\) is also an \((n'_*, \delta'_*, \tau'_*)\)-KHM.

In Lemma 35, adapted from [19, Lemma 8.2], we show that if an HMM satisfies Condition K, then it is also a KHM for some \((n_*, \delta_*, \tau_*)\). This is because Condition K ensures the existence of a sequence that induces a Birkhoff contraction coefficient less than 1 (a “good” sequence). However, the HMM may very well have many “good” sequences, possibly shorter. Thus, a given HMM that satisfies Condition K may be an \((n_*, \delta_*, \tau_*)\)-KHM for many different combinations of \(n_*, \delta_*, \tau_*\). Since the bounds we develop are dependent on the values of \(n_*, \delta_*, \tau_*\), it is worthwhile to seek the combination that yields the best bound.

**Lemma 35.** If the HMM \((A_n, B_n)\) satisfies Condition K then there exist a positive integer \(n_*\) and constants \(\delta_* < 1\) and \(0 \leq \tau_* < 1\) such that \((96)\) is satisfied.

**Proof:** By Condition K there exist positive integers \(k_0, l_0\) and numbers \(\gamma_0 > 0\) and \(0 \leq \tau_* < 1\) such that

1. For any \(i, j \in A_\text{.} M_{i,j}^{n_0} \geq \gamma_0\). This follows from \(M\) being aperiodic and irreducible, so some power of it must be strictly positive \([34, \text{Theorem 1.4}]\).
2. For some sequence \(\beta_{1}^{n_0}\) of elements of \(B\), the matrix \(M(\beta_{1}^{n_0})\) is nonzero and subrectangular. Existence of such sequences is guaranteed by Condition K. We denote \(\tau_* = \tau(M(\beta_{1}^{n_0}))\). Since \(M(\beta_{1}^{n_0})\) is subrectangular, \(0 \leq \tau_* < 1\).

Denote by \(A'\) the set of states that can lead to \(f^{-1}(\beta_1)\),

\[
A' = \left\{ a \in A \mid \|e_a^TM(\beta_{1}^{n_0})\|_1 > 0 \right\}.
\]

That is, there is positive probability that the next observation after any state in \(A'\) is the first observation \(\beta_1\) of the word \(\beta_{1}^{n_0}\). Since Condition K is satisfied, \(A'\) is not empty, so that

\[
a_0 = \min_{a \in A'}\|e_a^TM(\beta_{1}^{n_0})\|_1 > 0.
\]

We claim that \((96)\) is satisfied with \(n_* = k_0 + l_0\) and \(\delta_* = 1 - a_0\gamma_0 < 1\). Indeed, for any \(a_0 \in A\),

\[
P\left(\tau(M(B_{1}^{n_0})) \leq \tau_* \mid A_0 = a_0\right)
\]

\[
\geq P\left(\tau(M(B_{k_0+l_0}^{n_0})) \leq \tau_* \mid A_0 = a_0\right)
\]

\[
\geq \sum_{a \in A'} P\left(B_{k_0+l_0}^{n_0} = \beta_{1}^{n_0} \mid A_0 = a_0\right)
\]

\[
= \sum_{a \in A} P\left(B_{k_0+l_0}^{n_0} = \beta_{1}^{n_0} \mid A_0 = a_0\right)
\]

\[
= \sum_{a \in A} P\left(B_{k_0+l_0}^{n_0} = \beta_{1}^{n_0} \mid A_0 = a_0\right) \cdot P\left(A_{k_0} = a_0\right)
\]

\[
\geq 0 \gamma_0,
\]

where (a) is by Corollary 25 and Lemma 29, (b) is by Condition K, (c) is by the Markov property, and (d) is by \((96)\).

Let us now define the random variables \(N_k(\tau), k \geq 1\), by

\[
N_1(\tau) = \min\{n : \tau(M(B_{1}^{n})) \leq \tau\},
\]

\[
N_{k+1}(\tau) = \min\{n : \tau(M(B_{N_k+1}^{n})) \leq \tau\}, \quad k \geq 1.
\]

That is, the random variable \(N_1(\tau)\) is the time of the first occurrence of a sequence that induces a Birkhoff contraction coefficient \(\tau\) or less. In other words, \(N_1(\tau)\) is the smallest value of \(n\) such that \(M(B_{1}^{n})\) is a subrectangular matrix with Birkhoff contraction coefficient \(\tau\) or less. Similarly, the random variable \(N_k(\tau)\) is the gap between the \((k-1)\)th and \(k\)th occurrences of such sequences.

The following lemma and corollary are adapted from [19, Lemma 8.3], which was stated in [19] without proof.

**Lemma 36.** Let \((A_n, B_n)\) be an \((n_*, \delta_*, \tau_*)\)-KHM. If \(\delta_* > 0\), there exist \(\gamma > 0\) and \(0 \leq \rho < 1\) so that for any positive integer \(n_1\),

\[
P\left(N_1(\tau_*) \geq n_1 \mid A_0 = a_0\right) \leq \gamma \rho^{n_1}, \quad \forall a_0 \in A.
\]

**Proof:** Let \(T_1 = 1\) and denote, for any positive integer \(k\), the random variable \(T_k = \tau(M(B_{1}^{n})\))]. Observe that, by \((96)\),

\[
P(T_1 \leq \tau_* | A_0 = a_0) \geq 1 - \delta_* \quad \forall a_0 \in A.
\]

We now show that for any positive integer \(k\), and any \(a_0 \in A\),

\[
P(T_k > \tau_* | T_{k-1} > \tau_*; A_0 = a_0) \leq \delta_*.
\]

We will demonstrate this for \(k = 2\), as the proof for all other values of \(k\) is the same. For any \(a_0 \in A\),

\[
P(T_2 \leq \tau_* | T_1 > \tau_*; A_0 = a_0)
\]

\[
= P\left(\tau(M(B_{1}^{n})) \leq \tau_* \mid T_1 > \tau_*; A_0 = a_0\right)
\]

\[
\geq \sum_{a \in A} P\left(\tau(M(B_{1}^{n}) \leq \tau_* \mid T_1 > \tau_*; A_0 = a_0\right)
\]

\[
= \sum_{a \in A} P\left(\tau(M(B_{1}^{n}) \leq \tau_* \mid A_0 = a_0\right)
\]

\[
\leq \sum_{a \in A} P\left(T_1 > \tau_* \mid A_0 = a\right) p(a)
\]

\[
\leq 1 - \delta_*
\]

where (a) is because, by Lemma 29, if \(\tau(M(B_{1}^{n})) \leq \tau_*\) then \(\tau(M(B_{1}^{n})) \leq \tau_*\); in (b) we denoted \(p(a) = P(A_{n_*} = a | T_1 > \tau_*; A_0 = a_0); (c) is by the Markov property; and (d) is by \((96)\). Rearranging yields \((98)\). We remark that \((98)\) is also true without conditioning on \(\{T_{k-1} > \tau_*\}\).

Thus,

\[
P(T_k > \tau_* | A_0 = a_0)
\]

\[
= P(T_k > \tau_* | T_{k-1} > \tau_*; A_0 = a_0) \cdot P(T_{k-1} > \tau_*; A_0 = a_0).
\]

\[
\leq \delta_* P(T_{k-1} > \tau_*; A_0 = a_0)
\]

\[
\leq \delta_* P(T_{k-1} > \tau_*; A_0 = a_0).
\]
where (a) is by Lemma 29, by which the second summand in the first equality must be 0, and (b) is by (98). We conclude that for any integer $k$ and any $a_0 \in \mathcal{A}$,
\[ P\left( N_1(\tau_*) > k n_* | A_0 = a_0 \right) = P\left(T_k > \tau_* | A_0 = a_0 \right) \leq \delta_*^k. \]
Hence, for any positive integer $n_1$ (not necessarily a multiple of $n_*$) and any $a_0 \in \mathcal{A}$,
\[ P\left( N_1(\tau_*) \geq n_1 | A_0 = a_0 \right) \leq \delta_*^{n_1/n_* - 1}. \]
Rearranging, this yields
\[ P\left( N_1(\tau_*) \geq n_1 | A_0 = a_0 \right) \leq \frac{1}{\delta_*} \cdot \left( \delta_*^{1/n_*} \right)^{n_1}. \]
Thus, we obtain (97) with $\gamma = 1/\delta_*$ and $\rho = \delta_*^{1/n_*}$. To complete the proof, observe that $0 < \rho < 1$ since $0 < \delta_* < 1$.

We imposed $\delta_* > 0$ in Lemma 36 because this is the more interesting case. Clearly, Lemma 36 also holds when $\delta_* = 0$, albeit with different $\gamma, \rho$. However, we can do better in this case. Namely, if $\delta_* = 0$ for some $n_*$, this implies that at time $n_*$ the sequence of observations is ensured to induce Birkhoff contraction coefficient less than $\tau_*$. In this case, we can obtain a much simpler bound on the mutual information. We will return to this point in the proof of Theorem 33.

The upper bound in (97) is independent of $a_0$. Therefore, whenever $(A_n, B_n)$ is an $(n_*, \delta_*, \tau_*)$-KHHM and $\delta_* > 0$, we conclude that
\[ P\left( N_1(\tau_*) \geq n_1 \right) \leq \gamma \rho^{n_1}. \]
More generally, we have the following corollary.

**Corollary 37.** Let $(A_n, B_n)$ be an $(n_*, \delta_*, \tau_*)$-KHHM with $\delta_* > 0$. Then, there exist $\gamma > 0$ and $0 \leq \rho < 1$ such that for any positive integers $n_1, n_2, \ldots, n_m$,
\[ P\left( N_1(\tau_*) \geq n_1, N_2(\tau_*) \geq n_2, \ldots, N_m(\tau_*) \geq n_m \right) \leq \gamma^m \rho^{n_1 + n_2 + \cdots + n_m}. \quad (99) \]

**Proof:** For brevity, we denote $N_k = N_k(\tau_*)$. Since
\[ P\left( N_1 \geq n_1, N_2 \geq n_2, \ldots, N_m \geq n_m \right) = \prod_{k=1}^{m} P\left( N_k \geq n_k | N_i \geq n_i, i < k \right), \]
(99) will follow if $P\left( N_k \geq n_k | N_i \geq n_i, i < k \right) \leq \gamma \rho^{n_k}$. Indeed, for any $k$ we have
\[ P\left( N_k \geq n_k | N_i \geq n_i, i < k \right) = \sum_a P\left( N_k \geq n_k, A_{N_k-i} = a | N_i \geq n_i, i < k \right) \]
\[ = \sum_a P\left( N_k \geq n_k | A_{N_k-i} = a \right) P\left( A_{N_k-i} = a | N_i \geq n_i, i < k \right) \]
\[ = \gamma \rho^{n_k} \sum_a P\left( A_{N_k-i} = a | N_i \geq n_i, i < k \right) \]
\[ = \gamma \rho^{n_k}, \]
where (a) is by definition of $N_k$ and (b) is by (97).

**Proposition 38.** Let $(A_n, B_n)$ be an $(n_*, \delta_*, \tau_*)$-KHHM with $\delta_* > 0$. Denote
\[ \gamma = \frac{1}{\delta_*}, \quad \alpha = \gamma \cdot \log |\mathcal{A}|, \quad \rho = \delta_*^{1/n_*} < 1. \]
Then, for any $m \leq n$ we have
\[ I(A_0; A_{n+1} | B^m_n) \leq 4 \log \left( \frac{1 + \tau_*}{1 - \tau_*} \right) \tau_*^m + \alpha \frac{(\gamma n)^m}{m!} \rho^{m+1}. \quad (100) \]

**Proof:** Observe that the right-hand side of (99) depends only on the sum $n_1 + n_2 + \cdots + n_m$, and not the values of the individual values of $n_k$. Denote by $p(n, m)$ the number of positive integer $m$-tuples $(n_1, n_2, \ldots, n_m)$ such that $n = n_1 + n_2 + \cdots + n_m$, where each integer $n_k \geq 1$. In [42, p. 38], it is shown that $p(n, m) = \binom{n-1}{m-1}$. Thus, by (99),
\[ P\left( \sum_{k=1}^{m} N_k(\tau_*) \geq n \right) \leq p(n, m) \gamma^m \rho^n \]
\[ = \binom{n-1}{m-1} \gamma^m \rho^n \]
\[ \leq (n-1)^m \rho^n. \]

Next, consider the matrix product $M(B^m_n)$. We wish to count, in this product, the number of non-overlapping occurrences of contiguous sequences of matrices whose product has Birkhoff contraction coefficient at most $\tau_*$. This is accomplished by the integer-valued random variable
\[ D_n = D_n(\tau_*) = \max \left\{ m : \sum_{k=1}^{m} N_k(\tau_*) \leq n \right\}. \]

From the above discussion,
\[ P\left( D_n \leq m \right) = P\left( D_n < m + 1 \right) \]
\[ = P\left( \sum_{k=1}^{m+1} N_k(\tau_*) \geq n + 1 \right) \]
\[ \leq \gamma \frac{(\gamma n)^m}{m!} \rho^{m+1}. \quad (101) \]

Recall from (93) that $I(A_0; A_{n+1} | B^m_n) = \mathbb{E}[J]$, where we have denoted, for brevity,
\[ J \triangleq \log \left( \frac{\| e_{A_0}^T M(B^m_n)^T A_{n+1} \|_1}{\| \pi_T M(B^m_n)^T A_{n+1} \|_1} \cdot \frac{\| \pi_T^T M(B^m_n) \|_1}{\| e_{A_0}^T M(B^m_n) \|_1} \right). \]

This is a conditional mutual information. In particular, for any fixed sequence $b^m_1$ we have
\[ 0 \leq I(A_0; A_{n+1} | B^m_n = b^m_1) = \mathbb{E} \left[ J | B^m_n = b^m_1 \right] \leq \log |\mathcal{A}|, \quad (102) \]
where the inequalities are due to the properties of mutual information — it is nonnegative and upper-bounded by the logarithm of the alphabet size. The random variable $D_n$ is
a function of $B^n_i$ — given any realization $b^n_i$ of $B^n_i$, we can compute the value of $D_n$ precisely. For any $m \leq n$,

$$
E[J|D_n > m]P(D_n > m) = \sum_{b^n_i : D_n > m} E[J|B^n_i = b^n_i]P(B^n_i = b^n_i)
$$

(a) \leq \sum_{b^n_i : D_n \leq m} E[J|B^n_i = b^n_i]P(B^n_i = b^n_i)

(b) \leq 4 \log \left( \frac{1 + \tau_s}{1 - \tau_s} \right) \cdot \tau_s^m 

(103)

where (a) is because $D_n$ is integer valued and (b) is by Lemma 29 and Corollary 32. Moreover,

$$
E[J|D_n \leq m]P(D_n \leq m) = \sum_{b^n_i : D_n \leq m} E[J|B^n_i = b^n_i]P(B^n_i = b^n_i)
$$

(a) \leq |A| \cdot P(D_n \leq m)

(b) \leq \log |A| \cdot \frac{(ny)^m}{m!} \rho^{n+1} 

(104)

where (a) is by the right-hand inequality of (102) and (b) is by (101).

Thus, for any $m \leq n$ we have by (103) and (104),

$$
I(A_0; A_{n+1}|B^n_i) = E[J] = E[J|D_n > m]P(D_n > m) + E[J|D_n \leq m]P(D_n \leq m)
$$

\leq \log \left( \frac{1 + \tau_s}{1 - \tau_s} \right) \cdot \tau_s^m + (\gamma \cdot \log |A|) \cdot \frac{(ny)^m}{m!} \rho^{n+1}.

Denoting $\alpha = \gamma \cdot \log |A|$ completes the proof.

Remark 10. We note in passing that, if desired, one can set $m = \theta n$ in (100) and obtain an upper bound that vanishes with $n$, provided that $\theta$ is sufficiently small. To this end, we use the inequality $m! \geq (m/e)^m$, see [42, p. 52]. We set $m = \theta n$, and upper-bound the second summand in the right-hand side of (100) to obtain

$$
\alpha \cdot \left( \frac{ny}{m!} \rho^{n+1} \right) \leq \alpha \rho \cdot \left( \frac{\gamma e}{\theta} \right)^n.
$$

The right-hand side of the above inequality vanishes with $n$ for small enough $\theta$. To see this, observe that $\lim_{\theta \to 0} (\gamma e/\theta)^\theta = 1$,\(^{11}\) so we are ensured that if $\theta$ is small enough, $\rho \cdot (\gamma e/\theta)^\theta < 1$.

That said, taking $m = \theta n$ might not be the best strategy for minimizing $\theta$ in the right-hand side of (100). A different strategy is outlined in the proof of Theorem 33.

We are now ready to prove Theorem 33.

Proof of Theorem 33: By Lemma 35, $(A_n, B_n)$ is an $(n_*, \delta_*, \tau_*)$-KHMM for some $n_*, \delta_*, \tau_*$. Let

$$
m = \left\lfloor \log_{\tau_*} \left( \frac{\epsilon}{2 \cdot 4 \log \left( \frac{1 + \tau_s}{1 - \tau_s} \right)} \right) \right\rfloor.
$$

\(^{11}\)Indeed, since $(1/\theta)^\theta = e^{\theta \ln(1/\theta)}$ and by continuity of the exponential function at $0$, it suffices to show that $\lim_{\theta \to 0} \theta \ln(1/\theta) = 0$, This, in turn, holds by L'Hôpital's rule: $\lim_{\theta \to 0} \theta \ln(1/\theta) = \lim_{\theta \to 0} \ln(1/\theta)/(1/\theta) = \lim_{\theta \to 0} (-1)/(1/\theta) = \lim_{\theta \to 0} (-1/\theta) = 0$.

Case 1: If $\delta_s = 0$ then at time $\lambda = (m + 1)n_*$ the sequence $B^1_i$ can be divided into $m + 1$ contiguous sequences of length $n_*$, each inducing a Birkhoff contraction coefficient less than $\tau_*$. Therefore, using Corollary 32 we obtain that in this case for any $n \geq \lambda$,

$$
I(A_0; A_{n+1}|B^n_i) \leq 4 \log \left( \frac{1 + \tau_s}{1 - \tau_s} \right) \tau_s^m \leq \frac{\epsilon}{2}.
$$

Case 2: In the general case, $\delta_s > 0$ and we turn to Proposition 38. For $m$ fixed as above, we set $\lambda$ as the smallest integer greater than or equal to $m$ such that for any $n \geq \lambda$ we have

$$
\frac{\alpha(\gamma n)^m \rho^{n+1}}{m!} \leq \frac{\epsilon}{2},
$$

where $\gamma = 1/\delta_*, \alpha = \gamma \cdot \log |A|$, and $\rho = \delta_*/\lambda_\star$. Such $\lambda$ exists since $m$ is fixed and $\rho < 1$. For this $m$ and any $n \geq \lambda$, the right-hand side of (100) is upper-bounded by $\epsilon$. \(\blacksquare\)

Discussion. The upper bound in Proposition 38 is generally quite loose. We only count non-overlapping occurrences of “good” sequences, known to have Birkhoff contraction coefficient less than some $\tau_\star$, with lengths that are multiples of some $n_\star$. There may actually be many other subsequences — possibly shorter — that induce Birkhoff contraction coefficients less than 1, and we ignore those. Moreover, most occurrences of “good” sequences appear as the suffix of longer sequences. By Lemma 29, the induced Birkhoff contraction coefficient of these longer sequences will be smaller than that of the “good” sequences. Moreover, the values of $\gamma$ and $\rho$ are conservative.

A given KHMM may be associated with many combinations of $(n_\star, \delta_*, \tau_\star)$. Thus, one needs to carefully select the right combination of these parameters to minimize $\lambda$ in Theorem 33. A more refined analysis, that considers a KHMM for which multiple combinations $(n_\star, \delta_*, \tau_\star)$ are known may yield better bounds.

Nevertheless, even with this loose bound, we are able to ensure that the desired mutual information vanishes for sufficiently large $\lambda$. In practice, for a given process, the mutual information will be below the desired threshold much earlier than promised in Proposition 38.

Appendix A

Proof of Fast Polarization

In the fast stage of our construction, Arikan polar codes are designed based on recursive upper bounds on distribution parameters, such as the Bhattacharyya parameter. In this appendix we show that this procedure leads to fast polarization universally. Fast polarization results are usually of the flavor: “if the polar code length is large enough, then fast polarization is obtained.” This “large enough” length is related to the process for which the polar code is designed. In a universal setting, however, we must design the fast stage before knowing which process the code is to be used for. We show that it is indeed possible to determine this length regardless of the process. This is afforded because the slow stage is $(\eta, \mathcal{L}, 3\ell)$-monopolarizing.

Fast polarization is the phenomenon described in the following lemma. To keep the discussion focused, we present it for a special case of binary polar codes based on Arikan’s kernel.
Lemma 39 ([3], [6], [43]). Let $B_1, B_2, \ldots$ be independent and identically distributed random variables with $\mathbb{P}(B_i = 0) = \mathbb{P}(B_i = 1) = 1/2$. Let $Z_0, Z_1, \ldots$ be a $\{0, 1\}$-valued random process such that

$$Z_{n+1} \leq \kappa \cdot \begin{cases} 2Z_n^2, & B_{n+1} = 0, \\ Z_n, & B_{n+1} = 1, \end{cases}, \quad n \geq 0, \tag{105}$$

where $\kappa > 1$. If $Z_n$ converges almost surely to a $\{0, 1\}$-valued random variable $Z_\infty$ then for every $0 < \beta < 1/2$, we have

$$\lim_{n \to \infty} \mathbb{P}\left(Z_n \leq 2^{-2n\beta}\right) = \mathbb{P}(Z_\infty = 0). \tag{106}$$

Fast polarization was first stated and proved in [3]. It was later generalized by Şaşoğlu (see, e.g., [6, Lemma 4.2]). A simpler proof of a stronger result for the general case can be found in [43]. Our fast polarization result is based on the proof of [43].

For example, $Z_n$ might be the Bhattacharyya parameter of a randomly-selected polarized s/o-pair (tantamount to a synthetic channel, in a channel-coding setting), which is an upper-bound on the probability of error of estimating the symbol from its observation. In the memoryless case, the recursion (105) for the Bhattacharyya parameter with $\kappa = 2$ was established in [2, Proposition 5]. Under memory, (105) was shown in [13, Theorem 2], with $\kappa = 2\phi_0$, where $\phi_0$ is a mixing parameter of the process; mixing parameters are defined in Lemma 6. Thus, the Bhattacharyya parameter polarizes fast to 0 with or without memory.

The proof in [43] establishes (106) by showing that for every $\delta > 0$ there exists an $n_0$ such that

$$\mathbb{P}(Z_\infty = 0) - \delta \leq \mathbb{P}\left(\forall n \geq n_0, Z_n \leq 2^{-2n\beta}\right) \leq \mathbb{P}(Z_\infty = 0).$$

The magnitude of $n_0$ depends on two factors: the almost-sure convergence of $Z_n$ to $Z_\infty$ and the law of large numbers. The latter is independent of the process, but the former one is not. The proof utilizes the almost-sure convergence of $Z_n$ only for the following consequence. Recalling that $Z_n$ converges almost surely to a $\{0, 1\}$-valued random variable, for any $\epsilon_0, 0 > 0$ and $\delta_0 > 0$ there must be an $n_0$ such that

$$\mathbb{P}(Z_n \leq \epsilon_0) \geq \mathbb{P}(Z_\infty = 0) - \delta_0, \quad \forall n \geq n_0. \tag{107}$$

We reiterate that $n_0$ is process-dependent.

In our universal setting, the fast polarization stage occurs after the slow polarization stage. Specifically, it operates on s/o-pairs whose conditional entropy — and thus also Bhattacharyya parameter — is universally smaller than $\eta$, which can be set as small as desired.

The ability to set $\eta$ as small as desired is the key to obtaining universal fast polarization results. Namely, we prove the following lemma.

Lemma 40. Let $B_1, B_2, \ldots$ be independent and identically distributed random variables with $\mathbb{P}(B_i = 0) = \mathbb{P}(B_i = 1) = 1/2$. Let $Z_0, Z_1, \ldots$ be a $\{0, 1\}$-valued random process that satisfies (105) for some $\kappa > 1$. Fix $0 < \beta < 1/2$. Then, for every $\delta > 0$ there exist $\eta > 0$ and $n_0$ such that if $Z_0 \leq \eta$ then

$$\mathbb{P}\left(Z_n \leq 2^{-2n\beta}\right) \text{ for all } n \geq n_0 \geq 1 - \delta. \tag{108}$$

Crucially, $\eta$ and $n_0$ depend on the process $Z_n$ only through $\kappa$. Inspection of the proof of [43] reveals that Lemma 40 will be true once it is shown that for any $\epsilon_0, 0 > 0$ and $\delta_0 > 0$ there exists $n_0$ such that

$$\mathbb{P}(Z_n \leq \epsilon_0 \text{ for all } n \geq n_0) \geq 1 - \delta'.$$

The crux of our proof will be to show that we can set $\eta > 0$ and $n_0$ such that the above holds. We will need an auxiliary result, Corollary 42, which follows from Lemma 41, introduced and proved below.

Remark 11. Our statement of Lemma 40 is for a fast polarization stage based on Arikan’s kernel. This is done for the sake of simplicity. However, the lemma holds true for the more general case of other kernels. The key technical tool in the proof, Lemma 41, is stated in a general manner, enabling its use for other kernels without change.

Let $T_1, T_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables. Denote by $T$ a random variable distributed according to the same distribution as each of the random variables $T_i, i \in \mathbb{N}$. We assume that $T$ is bounded; in particular, there exist positive reals $a, b > 0$ such that

$$-b \leq T \leq a,$$

and for every $\epsilon > 0, \mathbb{P}(T > a - \epsilon) > 0$. We further assume that

$$\mu \triangleq \mathbb{E}[T] < 0. \tag{110}$$

We define the random walk

$$J_n = \sum_{i=1}^n T_i, \quad n \in \mathbb{N}.$$

For every $\alpha > 0$, define the events

$$A_\alpha(n) = \{J_m \geq \alpha \text{ for some } m \leq n\}$$

and

$$A_\alpha = \{J_m \geq \alpha \text{ for some } m \in \mathbb{N}\}.$$

14More generally, fast polarization of high-entropy indices may also be of interest, e.g., in source-coding applications. The universal stage also provides us with s/o-pairs whose conditional entropy is as close to 1 as desired. Due to forgetfulness (see the proof of Lemma 20, stopping short of the last inequality, (f)), this is true also when conditioning on the boundary states, by taking $L_0$ large enough. Under memory, fast polarization of high-entropy s/o-pairs is obtained through boundary-state-informed parameters, namely the total variation distance (see [14]). It was shown in [14, Proposition 12] that the boundary-state-informed total variation distance undergoes a recursion similar to (105). The required connections between the boundary-state-informed conditional entropy and the boundary-state-informed total variation distance can be found in [14, equation (4c)].
Observe that $A_\alpha(n) \subseteq A_\alpha(n+1)$ and $\bigcup_{n=1}^\infty A_\alpha(n) = A_\alpha$, so that by continuity of measure [37, Theorem 2.1],

$$
\mathbb{P}(A_\alpha) = \lim_{n \to \infty} \mathbb{P}(A_\alpha(n)).
$$

(111)

We denote by $A_\alpha^c$ the complementary event to $A_\alpha$. That is, $A_\alpha^c = \{ \alpha < a \text{ for all } n \in \mathbb{N} \}$. We then have the following lemma.

**Lemma 41.** There exists $r > 0$ such that for any $\alpha > 0$,

$$
\mathbb{P}(A_\alpha) \leq e^{-r\alpha}.
$$

(112)

Moreover, for any $0 < \gamma < 1$ and $n \in \mathbb{N}$,

$$
\mathbb{P}(J_n < n(1-\gamma)\mu) \leq 1 - e^{-2n((\gamma\mu)^2)}.
$$

(113)

Since $\mu < 0$ by (110) and $0 < \gamma < 1$ by assumption, then $n(1-\gamma)\mu < 0$ in (113). We will see in Corollary 42 below that Lemma 41 implies that for any negative threshold, there exists $n_\alpha \in \mathbb{N}$ and $\alpha > 0$ such that with probability arbitrarily close to 1, $J_n$ drops below that threshold for every $n \geq n_\alpha$ and never (for any $n \in \mathbb{N}$) visits above $\alpha$. This will be key to obtaining (109).

**Proof:** The proof combines two inequalities: (112) is essentially the Lundberg inequality [44, equation 15] and for (113) we call upon the Hoeffding inequality [45, Theorem 2]. Since the proof of the Lundberg inequality in [44] is for the continuous-time case, we provide a proof for the discrete-time case, adapted from the proof of [44].

Denote by $g(s)$ the moment-generating function of $T$. That is,

$$
g(s) = \mathbb{E}[e^{sT}].
$$

The expectation is well-defined as $e^{sT}$ is a non-negative random variable [37, equation 15.3]. Since $T$ is bounded by assumption, $g(s) < \infty$ for any $s \in \mathbb{R}$; hence, $g(s)$ is continuous over $\mathbb{R}$, see [46, Theorem 9.3.3]. Observe that $g(0) = 1$ and, by [37, equation 21.23] and (110), $g'(0) = \mathbb{E}[T] < 0$. Thus, $g(s)$ is decreasing at $s = 0$, so $g(s) < 1$ for $s$ small enough. On the other hand, by assumption on $T$,

$$
p \triangleq \mathbb{P}(T \geq a/2) = \mathbb{E}[\mathbb{1}\{T \geq a/2\}] > 0,
$$

where $\mathbb{1}\{}$ is an indicator random variable. Thus,

$$
g(s) \geq \mathbb{E}[e^{sT} \cdot \mathbb{1}\{T \geq a/2\}] \geq e^{s\alpha/2}p.
$$

In particular, if $s > (2/a)\ln(1/p)$, then $g(s) > 1$. Since $g(s)$ is continuous, there exists $s > 0$ such that $g(s) = 1$. Thus, we define

$$
r \triangleq \max_{s > 0} \{ s : \mathbb{E}[e^{sT}] = 1 \}.
$$

(114)

For the $r$ found above, denote

$$
\tilde{J}_n = e^{rT_n} = \prod_{i=1}^{n} e^{rT_i}.
$$

We claim that $\tilde{J}_n$, $n \in \mathbb{N}$, is a martingale. Indeed, since the $T_i$ are independent,

$$
\mathbb{E}[\tilde{J}_n \mid \tilde{J}_m, m < n] = \mathbb{E}[e^{rT_n} \cdot \tilde{J}_{n-1} \mid \tilde{J}_m, m < n]
$$

$$
= \tilde{J}_{n-1} \mathbb{E}[e^{rT_n}]
$$

$$
= \tilde{J}_{n-1}.
$$

where the last equality is by definition of $r$, (114). Define the (possibly infinite) stopping time

$$
\tau = \inf_n \{ n : J_n \geq \alpha \}.
$$

Then, by [47, Section 10.9], the stopped process

$$
\tilde{J}_{\tau \wedge n} \triangleq \begin{cases} 
J_n, & \tau > n, \\
J_\tau, & \tau \leq n
\end{cases}
$$

is also a martingale, and

$$
\mathbb{E}[\tilde{J}_{\tau \wedge n}] = \mathbb{E}[\tilde{J}_1] = 1.
$$

Observe that for any $n \in \mathbb{N}$, we have $\mathbb{P}(A_\alpha(n)) = \mathbb{P}(\tau \leq n)$. Thus,

$$
1 \geq \mathbb{E}[\tilde{J}_{\tau \wedge n}] = \mathbb{E}[\tilde{J}_\tau] = \mathbb{E}[\tilde{J}_n] \geq e^{r\alpha} \mathbb{P}(A_\alpha(n)).
$$

(115)

where (a) is because $\tilde{J}_{\tau \wedge n} \geq 0$, (b) is by definition of $\tau$ and of $\tilde{J}_{\tau \wedge n}$, and (c) is because $r > 0$ by definition. Rearranging, we obtain that for any $n \in \mathbb{N}$,

$$
\mathbb{P}(A_\alpha(n)) \leq e^{-r\alpha}.
$$

Thus, by (111),

$$
\mathbb{P}(A_\alpha) = \lim_{n \to \infty} \mathbb{P}(A_\alpha(n)) \leq e^{-r\alpha}.
$$

This completes the proof of (112).

To prove (113), recall that by the Hoeffding inequality [45, Theorem 2], for any $t > 0$ we have

$$
\mathbb{P}(J_n \geq n(\mu + t)) \leq e^{-2n((\frac{t}{\gamma\mu})^2)}.
$$

In particular, for any $0 < \gamma < 1$, we may choose $t = \gamma|\mu| = -\gamma\mu > 0$ to obtain

$$
\mathbb{P}(J_n < n(1-\gamma)\mu) = 1 - \mathbb{P}(J_n \geq n(\mu + |\mu|))
$$

$$
\geq 1 - e^{-2n((\frac{\mu}{\gamma\mu})^2)}.
$$

This completes the proof.

**Corollary 42.** Under the same setting as in Lemma 41, for any $n_\alpha \geq 0$, $\alpha > 0$, and $0 < \gamma < 1$ we have

$$
\mathbb{P}\left(\big\{ \forall n \geq n_\alpha, J_n < n_\alpha(1-\gamma)\mu \big\} \cap A_\alpha^c\right) \geq 1 - \left(1 - e^{-2n((\frac{\gamma\mu}{\gamma\mu})^2)}\right) \cdot e^{-2n_\alpha((\frac{\gamma\mu}{\gamma\mu})^2)} - e^{-r\alpha}.
$$

(115)
Proof: Note that
\[
\mathbb{P}(\forall n \geq n_a, J_n < n_a(1 - \gamma)\mu)
\]
\[
= \mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n_a(1 - \gamma)\mu\}\right)
\]
\[
\geq \mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n(1 - \gamma)\mu\}\right)
\]
\[
= 1 - \mathbb{P}\left(\bigcup_{n=n_a}^{\infty} \{J_n \geq n(1 - \gamma)\mu\}\right)
\]
\[
\geq 1 - \sum_{n=n_a}^{\infty} e^{-2n(\frac{\gamma}{\kappa})^2}
\]
\[
= 1 - \frac{1}{1 - e^{-2(\frac{\mu}{\kappa})^2}} \cdot e^{-2n_0(\frac{\mu}{\kappa})^2},
\]
(116)
where (a) is by (113) and the union bound. Observing that
\[
\mathbb{P}(\forall n \geq n_a, J_n < n(1 - \gamma)\mu)
\]
\[
= \mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n(1 - \gamma)\mu\} \cap A_n\right)
\]
\[
+ \mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n(1 - \gamma)\mu\} \cap A_n^c\right)
\]
\[
\leq \mathbb{P}(A_n) + \mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n(1 - \gamma)\mu\} \cap A_n^c\right),
\]
we obtain
\[
\mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n(1 - \gamma)\mu\} \cap A_n^c\right)
\]
\[
\geq \mathbb{P}(\forall n \geq n_a, J_n < n(1 - \gamma)\mu) - \mathbb{P}(A_n).
\]
Combining this inequality with (112) and (116) yields (115) and completes the proof.

Proof of Lemma 40: By inspection of the proof of [43], the lemma will be true once we show that for any \(\epsilon_a > 0\) and \(\delta' > 0\) there exist \(n_a\) and \(\eta\) such that if \(Z_0 > \eta\), then \(109\) holds. Thus, we fix \(\epsilon_a > 0\) and \(\delta' > 0\), and work toward this goal.

Let the process \(\tilde{Z}_0, \tilde{Z}_1, \ldots\) be defined as
\[
\tilde{Z}_0 = \ln Z_0,
\]
\[
\tilde{Z}_{n+1} = \begin{cases} 2\tilde{Z}_n + \ln \kappa, & B_{n+1} = 0, \\
\tilde{Z}_n + \ln \kappa, & B_{n+1} = 1, \end{cases} \quad n \geq 0.
\]
Then, by (105), \(\ln Z_n \leq \tilde{Z}_n\) for any \(n\). Therefore, (109) will be true once we show that there exists \(n_a\) and \(\eta\) such that if \(Z_0 = \eta\), then
\[
\mathbb{P}(\tilde{Z}_n \leq \ln \epsilon_a \text{ for all } n \geq n_a) \geq 1 - \delta'.
\]
Fix
\[
0 < \zeta < 1/\kappa^2
\]
such that \(\tilde{Z}_0 < \ln \zeta < 0\). Since \(\tilde{Z}_0 = \ln \eta\) by assumption, and since we may set \(\eta\) as small as desired, we can ensure that this is possible. We then have, by (105),
\[
\tilde{Z}_1 \leq \begin{cases} \tilde{Z}_0 + \ln \kappa + \ln \zeta, & B_0 = 0, \\
\tilde{Z}_0 + \ln \kappa, & B_0 = 1. \end{cases}
\]
If, further, \(\tilde{Z}_1 < \ln \zeta\) then the above inequality holds when \(Z_1\) and \(Z_0\) are replaced with \(Z_2\) and \(Z_1\), respectively. More generally, we define the process \(J_n, n \in \mathbb{N}\), by
\[
J_0 = Z_0 = \ln \eta,
\]
\[
J_{n+1} = J_n + T_{n+1}, \quad n \geq 0,
\]
where
\[
T_n = \begin{cases} \ln \kappa + \ln \zeta, & B_n = 0, \\
\ln \kappa, & B_n = 1, \end{cases} \quad n \geq 1.
\]
If \(J_i < \ln \zeta\) for all \(i \leq n\), then \(Z_n \leq J_n\).

Recall that \(B_1, B_2, \ldots\) is a sequence of i.i.d. random variables with \(\mathbb{P}(B_i = 0) = \mathbb{P}(B_i = 1) = 1/2\) for any \(i\). Thus, \(T_1, T_2, \ldots\) is a sequence of i.i.d. random variables. Denoting by \(T\) a random variable distributed according to their common distribution, we have \(\mathbb{P}(T = \ln \kappa) = \mathbb{P}(T = \ln \kappa + \ln \zeta) = 1/2\). In particular, \(T\) is bounded:
\[
-\ln \left(\frac{1}{\kappa}\right) = -b \leq T \leq a = \ln \kappa.
\]
Both \(a\) and \(b\) are positive by (118) and since \(\kappa > 1\) by assumption. By definition, for any \(\epsilon > 0\), \(\mathbb{P}(T > a - \epsilon) \geq \mathbb{P}(T = a) = 1/2\). Moreover, by (118),
\[
\mu = \mathbb{E}[T] = \frac{1}{2} \ln(\kappa^2 \zeta) < 0.
\]
Consequently, Corollary 42 holds for the random walk \(J_n - J_0 = \sum_{i=1}^{n} T_i, n \in \mathbb{N}\).

Let \(r > 0\) be the largest positive solution of the equation
\[
\mathbb{E}[e^{rT}] = \frac{(\kappa \gamma)^{r} + \kappa^{r}}{2} = 1.
\]
(119)
Such \(r\) exists, as shown in the proof of Lemma 41. Denote for brevity
\[
\theta \triangleq \frac{\mu}{a + b}.
\]
By Corollary 42, for any \(0 < \gamma < 1\) and \(n_a \geq 0\) we have
\[
\mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n - J_0 < n_a(1 - \gamma)\mu\} \cap A_{J_0+\ln \zeta}\right)
\]
\[
\geq 1 - \left(1 - e^{-2\gamma^2 \theta^2}\right)^{-1} e^{-2n_a \gamma^2 \theta^2} - e^{-r(-J_0+\ln \zeta)},
\]
(120)
where (a) is because \(\mu < 0\).

Observe that since \(J_n = J_0 + \sum_{i=1}^{n} T_i\) we have
\[
A_{J_0+\ln \zeta} = \left\{ \sum_{i=1}^{n} T_i < -J_0 + \ln \zeta \text{ for all } n \in \mathbb{N} \right\} = \{J_n < \ln \zeta \text{ for all } n \in \mathbb{N} \}.
\]
Consequently, under the event \(A_{J_0+\ln \zeta}\), we have \(\tilde{Z}_n \leq J_n\) for any \(n\). Hence,
\[
\mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n_a(1 - \gamma)\mu\} \cap A_{J_0+\ln \zeta}\right)
\]
lower-bounds the probability that \(\tilde{Z}_n \leq J_0 + n_a(1 - \gamma)\mu\) for all \(n \geq n_a\).

Recall that \(J_0 = \tilde{Z}_0 = \ln \eta\). It remains to set \(\eta\) and \(n_a\) such that \(\ln \eta < \ln \zeta, J_0 - n_a(1 - \gamma)\mu \leq \ln \epsilon_a\), and the right-hand side of (120) exceeds \(1 - \delta'\). Below we show one selection of
We set \( n \) to be such that \( |\mu| = (\ln 2)/2 \) and \( \theta = \ln 2/(2\ln(2K^*)) \). Further, our plan is to split \( \delta' \) equally among the two subtracted terms on the right-hand side of (120). We stress that these are arbitrary choices, and in practice should be optimized. We plug \( \delta' \) into (119) and compute \( r \), the largest possible solution of \( \kappa^*(2K^*)^r = 2 \).

Next, we set \( J_0 \) so that \( e^{-\gamma(J_0+\ln \zeta)} \leq \delta'/2 \); one choice is \( J_0 = \ln \zeta + \frac{1}{r} \ln(\delta'/2) \). Observe that indeed \( J_0 = \ln \eta < \ln \zeta \) since \( \delta' \geq 1 \). We thus take

\[
\eta = e^{\gamma} = \frac{1}{2\kappa^*} \left( \frac{\delta' \ln 2}{r} \right).
\]

We set \( n_d \) large enough such that both \( J_0 - n_d |\mu|/2 \leq \ln \epsilon_a \) and \( (1 - e^{-2\gamma \delta^2}) - 1 \leq \epsilon_a \) hold. That is, \( n_d = [n_d'] \), where

\[
n_d' = \left\lfloor \frac{\ln 2}{\ln \left( \frac{2}{\delta'} \cdot (1 - e^{-2\gamma \delta^2}) \right)} \right\rfloor.
\]

For the above \( \eta \) and \( n_d \), \( \mathbb{P}(Z_n \leq \ln \epsilon_a) \) for all \( n \geq n_d \) is at least \( 1 - \delta' \). Thus, (109) holds, and the proof is complete.

The parameters \( n_d \) and \( \eta \) found in the above proof depend on the process \( Z_n \) only through \( \kappa \). Thus, they universally apply to any process for which (105) holds. In particular, one can set \( \kappa \) in advance a universal length \( \bar{N} \) for the polar code in the fast stage.

The values of \( n_d \) and \( \eta \) are not optimized in the above proof, and the actual required length of the fast stage is expected to be shorter in practice. When designing a universal polar code, one can try out several small values of \( \eta \) and numerically run the recursion (117) until \( Z_n \) is sufficiently small for most indices. The above proof implies that if \( \eta \) is small enough and we run the recursion for sufficiently long, we are ensured that most indices will polarize fast.

### Appendix B

**Auxiliary Proofs for Section V-A**

We denote \( T_j = (X_j, Y_j), j \in \mathbb{Z}, \) with realization \( t_j \), and \( T^N_M = (X^N_M, Y^N_M) \) with realization \( \mathcal{T}^N_M \). For brevity, we denote \( P_{\mathcal{T}^N_M} = P_{\mathcal{T}^N_M}(t^N_M) \), and similarly \( P_{\mathcal{S}^N_M} = P_{\mathcal{S}^N_M}(s^N_M) \).

**Proof of Lemma 6**: Although (33a) was already proved in [14, Lemma 5], we provide a proof here for completeness.

We will prove that (33) holds with

\[
\psi_k = \begin{cases} 
\max_{s, \sigma} \frac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)}, & k > 0, \\
1, & k = 0,
\end{cases}
\]

and

\[
\phi_k = \min_{s, \sigma} \frac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)}, \quad k > 0, \\
0, & k = 0.
\]

Recall that by stationarity, \( P_{S_0} = P_{S_k} \) for any \( k \). Further, observe that by Bayes’ law,

\[
\frac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)} = \frac{\mathbb{P}(S_k = \sigma | S_0 = s)}{\mathbb{P}(S_k = \sigma)}.
\]

**Fig. 11.** Two blocks of a FAIM process, not necessarily of the same length. The state \( S_0 \), just before the first block, assumes value \( a \in \mathbb{S} \). The final state of the first block, \( S_L \), assumes value \( b \in \mathbb{S} \). The state \( S_M \), just before the second block, assumes value \( c \in \mathbb{S} \). The final state of the second block, \( S_N \), assumes value \( d \in \mathbb{S} \).

To prove (33), we first consider the case \( M > L \). Denote by \( a, b, c, d \) the values of states \( S_0, S_L, S_M, \) and \( S_N \), respectively (see Fig. 11). Then,

\[
P_{T^L_1, T^N_M} = \sum_{a, b, c, d} P_{T^L_1, T^N_M}(a, b, c, d) \]

where (a) follows from the definition of \( \psi_k \). This shows (33a). To see (33b) we follow the exact steps above up to just before inequality (a), and proceed with

\[
P_{T^L_1, T^N_M} \geq \phi_k \sum_{a \in \mathbb{S}_L} P_{T^L_1, T^N_M}(a, b, c, d) \sum_{b \in \mathbb{S}_L} P_{T^L_1, S_L}(b, d) P_{S_L} \]

Again, the inequality follows from the definition of \( \phi_k \).

For the case \( M = L \), we need only establish (33a), as (33b) is trivially true for \( M = L \). Again, \( a \) and \( d \) represent the values of states \( S_0 \) and \( S_N \). Both \( b \) and \( b' \) represent values of state \( S_L \); this distinction is to distinguish the summation variables of two different sums over values of \( S_L \). Thus,

\[
P_{T^L_1, T^N_M} = \sum_{a \in \mathbb{S}_L} P_{T^L_1, S_L}(a, b, c, d) P_{S_L} \sum_{b \in \mathbb{S}_L} P_{T^L_1, S_L}(b, d) P_{S_L} \]

where \( \psi_0 \) follows from the definition of \( \psi_0 \). This shows (33a). To see (33b) we follow the exact steps above up to just before inequality (a), and proceed with

\[
P_{T^L_1, T^N_M} \geq \phi_k \sum_{a \in \mathbb{S}_L} P_{T^L_1, S_L}(a, b, c, d) \sum_{b \in \mathbb{S}_L} P_{T^L_1, S_L}(b, d) P_{S_L} \]

Again, the inequality follows from the definition of \( \phi_k \).
where the inequality is by the definition of \( \psi_0 \) and because
\[
P_{T_l^{S_k|S_0}} \leq \sum_{s} P_{T_l^{S_k|S_0}},
\]
To see that that \( \psi_k \) is nonincreasing, observe that for any \( s, \sigma \in \mathcal{S} \):
\[
P_{S_{k+1},S_0|S_k}(\sigma, s) = \sum_{a \in \mathcal{S}} P_{S_{k+1}|S_k}(\sigma|a) \cdot P_{S_k}(a, s)
\]
\[
\leq \psi_k \sum_{a \in \mathcal{S}} P_{S_{k+1}|S_k}(\sigma|a) \cdot P_{S_k}(a) = \psi_k P_{S_{k+1}(\sigma)}P_{S_k}(s).
\]
Therefore, we must have \( \psi_{k+1} \leq \psi_k \). The proof that \( \phi_k \) is nondecreasing is similar, with \( \leq \psi_k \) replaced with \( \geq \phi_k \).
Finally, the asymptotic properties of \( \phi_k \) and \( \psi_k \) are due to \( S_j \) being an aperiodic and irreducible stationary finite-state Markov chain. For in this case there exist \( \gamma < 1 \) and \( 0 < \alpha < \infty \) such that for any \( s, \sigma \in \mathcal{S} \) and \( k \geq 0 \),
\[
|P_{S_k|S_0}(\sigma|s) - P_{S_k}(\sigma)| \leq \alpha \cdot \gamma^k,
\]
see [32, Theorem 4.3] for a proof. Rearranging and observing that \( \psi_0 < \infty \), we obtain that
\[
\frac{|P(S_0 = s, S_k = \sigma)|}{P(S_0 = s) P(S_k = \sigma)} - 1 \leq \psi_0 \cdot \alpha \cdot \gamma^k \xrightarrow{k \to \infty} 0.
\]
Hence, both \( \psi_k \) and \( \phi_k \) must tend to 1 exponentially fast as \( k \to \infty \).

**Proof of Lemma 7:** The FAIM process is forgetful, so we let \( \lambda \) be the \( \epsilon \)-recollection of the process. For this \( \lambda \), (34) is satisfied,

\[
I(S_0; S_{-k}, S_k|X_{-\ell}, Y_{-\ell}^m) = I(S_0; S_k|X_{-\ell}, Y_{-\ell}^m, S_k). \tag{121}
\]

We will upper-bound each of the terms on the right-hand side of (121) by \( \epsilon \), yielding the desired result.

For any \( m, \ell, k \) such that \( \min\{m, \ell\} \geq k \geq \lambda \) we have
\[
\epsilon \geq I(S_0; S_k|Y_{0}^k),
\]
\[
\geq I(S_0; (S_k, Y_{-\ell}^m)|Y_{0}^k)
\]
\[
= I(S_0; Y_{-\ell}^m|Y_{0}^k) + I(S_0; S_k|Y_{0}^m)
\]
\[
\geq I(S_0; S_k|Y_{0}^m)
\]
\[
\geq I((S_0, X_{-\ell}^m, Y_{-\ell}^m); S_k|Y_{0}^m)
\]
\[
= I(X_{-\ell}^m, Y_{-\ell}^m; S_k|Y_{0}^m) + I(S_0; S_k|X_{-\ell}^m, Y_{-\ell}^m)
\]
\[
\geq I(S_0; S_k|X_{-\ell}^m, Y_{-\ell}^m).
\]

We now justify the inequalities:

- (a) is by (34b) and stationarity.
- (b) is by (2), noting that (32) implies

\[
S_0 \not\rightarrow (S_k, Y_{-\ell}^k) \not\rightarrow (S_k, Y_{-\ell}^m);
\]
- (c) is because mutual information is nonnegative;
- (d) is by (2), noting that (32) implies

\[
S_k \not\rightarrow (S_0, Y_{0}^m) \not\rightarrow (S_0, X_{-\ell}^m, Y_{-\ell}^m)
\]

(observe that \( X_{-\ell}^m, Y_{-\ell}^m \) is “in the past” whereas \( Y_{0}^m \) is “in the future,” and the state \( S_0 \) is in between);
- (e) is because mutual information is nonnegative.

The derivation for the second term in the right-hand side of (121) is similar. For any \( m, \ell, k \) such that \( \min\{m, \ell\} \geq k \geq \lambda \) we have
\[
\epsilon \geq I(S_0; S_{-k}|X_{-\ell}^m, Y_{-\ell}^m)
\]
\[
\geq I((S_0, Y_{0}^m, S_k); S_{-k}|X_{-\ell}^m, Y_{-\ell}^m)
\]
\[
\geq I(S_0; S_{-k}|X_{-\ell}^m, Y_{-\ell}^m).
\]

Again, we justify the inequalities:

- (a) is by (35a) and stationarity.
- (b) is by (2), noting that (32) implies

\[
S_0 \not\rightarrow (S_{-k}, X_{-\ell}^m, Y_{-\ell}^m) \not\rightarrow (S_{-k}, X_{-\ell}^{m-1}, Y_{-\ell}^{m-1});
\]
- (c) is by the chain rule for mutual information
- (d) is by (2), noting that (32) implies

\[
S_{-k} \not\rightarrow (S_0, X_{-\ell}^m, Y_{-\ell}^m) \not\rightarrow (S_0, Y_{0}^m, S_k);
\]
- (e) is by the chain rule for mutual information.

This completes the proof.

**Proof of Corollary 8:** The FAIM process is forgetful, so we set \( \lambda \) as the \( \epsilon \)-recollection of the process. The corollary holds for \( k = 1 \) by Lemma 7. We proceed by induction. Assume that the corollary holds for \( k - 1 \geq 1 \), and we will show it holds for \( k \).

Let
\[
i' = [i_1 \ i_2 \ \cdots \ i_{k-1}],
\]
\[
i = [i_1 \ i_2 \ \cdots \ i_{k-1} \ i_k] = [i' \ i_k].
\]

For brevity, denote
\[
C_i = (X_i^{i+1}, \ Y_i^{i+1}).
\]

Our goal is thus to show that
\[
I(S_i; S_{i-L_0}, S_{i+L_0}, C_i) = I(S_i; S_{i-L_0}, S_{i+L_0}, S_{i-L_0}, S_{i+L_0}, S_{i-L_0}, S_{i+L_0}, C_i, C_i) \leq k \cdot 2\epsilon.
\]

Indeed,
\[
I(S_i; S_{i-L_0}, S_{i+L_0}, C_i) = I(S_i; S_{i-L_0}, S_{i+L_0}, S_{i-L_0}, S_{i+L_0}, C_i, C_i) \leq k \cdot 2\epsilon,
\]

where (a) is by the chain rule; (b) is by (2) and (32), used for the Markov chains (See Figure 12 for an illustration):

\[ S_F \leadsto (S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k}) \leadsto (S_{i_k-L_0}, S_{i_k+L_0}, S_{i_k+L_0}, S_{i_k+L_0}, C_{i_k}) \]

which holds because \( i_{k-1} \leq i_k - L_0 \leq i_k - L_0 \) so \((S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k})\) are independent of \(S_F\) given \(S_{i_k-L_0}, L_0\), which is part of \(S_{i_k+L_0}\), and

\[
S_{i_k} \leadsto (S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k}) \\
\leadsto (S_{i_k-L_0}, S_{i_k+L_0}, S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k})
\]

which again holds because \( i_{k-1} + L_0 \leq i_k - L_0 \leq i_k \), so \((S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k})\) are independent of \(S_F\) given \(S_{i_k-L_0}, L_0\); finally, (c) is because \(I(S_F; S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k}) \leq (k-1) \cdot 2e\) by the induction hypothesis and \(I(S_{i_k}; S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k}) \leq 2e\) by Lemma 7. This completes the proof.

### Appendix C

**Auxiliary Proofs for Section IV-B**

Recall from (1) that the binary entropy function \( h_2 : [0, 1] \rightarrow [0, 1] \) is defined by

\[
h_2(x) = -x \log x - (1-x) \log (1-x).
\]

This is a concave-\(\cap\) function that satisfies \( h_2(x) = h_2(1-x) \) for any \( x \in [0, 1] \), and it is monotone increasing over \([0, 1/2]\). The inverse of the binary entropy function is \( h_2^{-1} : [0, 1] \rightarrow [0, 1/2] \). The following three technical lemmas will be used to prove Lemma 13.

**Lemma 43.** For any \( 0 \leq x \leq 1/2 \),

\[
1 - h_2(x) \geq \frac{2}{\ln 2} \left( \frac{1}{2} - x \right)^2.
\]

**Proof:** Denote \( g(x) = 1 - h_2(x) \). Clearly, \( 1 = g(0) > 1/(2 \ln 2) \approx 0.721 \). For any \( \epsilon > 0 \), the function \( g(x) \) is 4 times continuously differentiable over \([\epsilon, 1/2]\). Therefore, by Taylor’s formula with remainder [48, Theorem 5.19], for any \( x \in [\epsilon, 1/2]\), there exists \( y \in [x, 1/2]\) such that

\[
g(x) = \frac{2}{\ln 2} \left( \frac{1}{2} - x \right)^2 + \frac{g^{(4)}(y)}{4!} \left( \frac{1}{2} - x \right)^4.
\]

However, \( g^{(4)}(y) = 2(y^{-3} + (1-y)^{-3})/\ln 2 > 0 \) for any \( y \in [\epsilon, 1/2] \). Hence, \( 1 - h_2(x) \geq 2(1/2 - x)^2/(\ln 2) \) for any \( 0 \leq x \leq 1/2 \) as well.

**Lemma 44.** For any \( 0 \leq y \leq x \leq 1/2 \),

\[
h_2(x) - h_2(y) \geq \frac{1}{\ln 2} (x - y) (1 - 2y).
\]

**Proof:** There is nothing to prove if \( x = y \), so we assume that \( y < x \). Due to the concavity of \( h_2(x) \), for any \( x_1 \leq x_2 \leq x_3 \) we have

\[
(x_3 - x_1)(h_2(x_2) - h_2(x_1)) \geq (x_2 - x_1)(h_2(x_3) - h_2(x_1)).
\]

(see, for example, [49, Section 1.4.3], or [50, Exercise 6.17]). Setting \( x_1 = y, x_2 = x, x_3 = 1/2 \) in (124) we obtain

\[
\left( 1 - \frac{1}{2} \right) (h_2(x) - h_2(y)) \geq (x - y) (1 - h_2(y)).
\]

Since \( y < x \leq 1/2 \) by assumption, \( 1/2 - y > 0 \). Therefore, we rearrange the above inequality and obtain

\[
h_2(x) - h_2(y) \geq (x - y) \frac{1 - h_2(y)}{1/2 - y} \geq \frac{1}{\ln 2} (x - y) (1 - 2y),
\]

where the rightmost inequality is by (122).

**Lemma 45.** For any \( x, y \in (0, 1/2) \), the function

\[
f(x, y) = h_2(h_2^{-1}(x) * h_2^{-1}(y)) - y
\]

is increasing in \( x \) and decreasing in \( y \).

**Proof:** Denote, for \( x, y \in (0, 1/2) \),

\[
g(x, y) = h_2(x * y) - h_2(y).
\]

Then, \( f(x, y) = g(h_2^{-1}(x), h_2^{-1}(y)) \). The function \( h_2(x) \) is monotone increasing over \([0, 1/2]\), so \( h_2^{-1}(x) \) is also monotone increasing over \([0, 1/2]\). Therefore, the claim will be true once we establish that \( g(x, y) \) is increasing in \( x \) and decreasing in \( y \).

To this end, recall the function

\[
\arctanh(x) = \frac{1}{2} \ln \left( \frac{1 + x}{1 - x} \right),
\]

defined for \( x \in [0, 1] \). This is an increasing function of \( x \) (since its derivative is \( 1 - x^2)^{-1} \), which is positive). Moreover, \( \arctanh(x) > 0 \) for \( x > 0 \).

Now,

\[
\frac{\partial g(x, y)}{\partial x} = \frac{2}{\ln 2} (1 - 2y) \arctanh ((1 - 2x)(1 - 2y)).
\]

This is positive since \( \arctanh(z) > 0 \) for \( z > 0 \), and both \( (1 - 2x) > 0 \) and \( (1 - 2y) > 0 \). Thus, \( g(x, y) \), and by proxy \( f(x, y) \), is increasing in \( x \). Next,

\[
\frac{\partial g(x, y)}{\partial y} = \frac{2}{\ln 2} \left( (1 - 2x) \arctanh ((1 - 2x)(1 - 2y)) - \arctanh(1 - 2y) \right)
\]

\[
\leq \frac{2}{\ln 2} \left( (1 - 2x) \arctanh(1 - 2y) - \arctanh(1 - 2y) \right)
\]

\[
= \frac{2}{\ln 2} \left( (1 - 2x) - 1 \right) \cdot \arctanh(1 - 2y)
\]

\[
< 0
\]

where the first inequality is because \( (1 - 2x)(1 - 2y) < (1 - 2y) \) and \( \arctanh() \) is increasing. Thus, \( g(x, y) \), and by proxy \( f(x, y) \), is decreasing in \( y \).

**Proof of Lemma 13:** It was shown in [6, Lemma 2.1] that

\[
\sum_{a,b} p_a q_b h_2(\alpha_a \ast \beta_b) \geq h_2(h_2^{-1}(A) \ast h_2^{-1}(B)),
\]
where
\[ A = \sum_a p_a h_2(\alpha_a), \quad B = \sum_b q_b h_2(\beta_b). \]

Therefore,
\[ \sum_{a,b} p_{ab} (h_2(\alpha_a \beta_b) - h_2(\beta_b)) \geq h_2(h_2^{-1}(A) + h_2^{-1}(B)) - B = f(A,B), \]
where \( f(\cdot,\cdot) \) was defined in (125). By (54), \( A \geq \xi_1 \) and \( B \leq \xi_2 \).

Since, by Lemma 45, \( f(A,B) \) is increasing in \( A \) and decreasing in \( B \), we conclude that
\[ \sum_{a,b} p_{ab} (h_2(\alpha_a \beta_b) - h_2(\beta_b)) \geq h_2(h_2^{-1}(\xi_1) + h_2^{-1}(\xi_2)) - \xi_2. \]

Define, therefore,
\[ \Delta(\xi_1,\xi_2) = h_2(h_2^{-1}(\xi_1) + h_2^{-1}(\xi_2)) - \xi_2. \] (126)

It remains to show that \( \Delta(\xi_1,\xi_2) > 0 \).

To this end, observe that for any \( x, y \in (0,1/2) \),
\[ h_2(x + y) - h_2(y) \geq \frac{1}{\ln 2} (x + y - y) \cdot (1 - 2y) = \frac{1}{\ln 2} x(1 - 2y)^2. \]

where (a) is by (123). Therefore,
\[ \Delta(\xi_1,\xi_2) \geq \frac{1}{\ln 2} h_2^{-1}(\xi_1) \left(1 - 2h_2^{-1}(\xi_2)\right)^2 > 0. \]

We note in passing that the expression for \( \Delta(\xi_1,\xi_2) \) derived here (or its lower bound) may be used to obtain a tighter lower bound than that of [13, Lemma 11].

**APPENDIX D**

**AUXILIARY PROOFS FOR SECTION V-C**

**Proof of Lemma 22.** Denote \( F = f(A), \tilde{F} = f(\tilde{A}), G = g(A), \) and \( \tilde{G} = g(\tilde{A}). \) For any \( f_0 \in \{0,1\} \), \( g_0 \in \mathcal{G} \), we abuse notation and write
\[ p(f_0, g_0) = \mathbb{P}(F = f_0, G = g_0) = \sum_{a : f(a) = f_0} p(a), \] (127a)
\[ q(f_0, g_0) = \mathbb{P}(\tilde{F} = f_0, \tilde{G} = g_0) = \sum_{a : f(a) = f_0} q(a). \] (127b)

With this notation we also have \( p(g_0) = \sum_{f_0} p(f_0, g_0) = \mathbb{P}(G = g_0) \) and \( p(f_0 | g_0) = \mathbb{P}(F = f_0 | G = g_0) \). The distributions \( q(g_0) \), \( q(f_0 | g_0) \) are similarly defined. By (68) and (127) we have for all \( f_0 \in \{0,1\} \) and \( g_0 \in \mathcal{G} \),
\[ (1 - \epsilon) q(f_0, g_0) \leq p(f_0, g_0) \leq (1 + \epsilon) q(f_0, g_0), \]
\[ (1 - \epsilon) q(g_0) \leq p(g_0) \leq (1 + \epsilon) q(g_0). \] (128)

Therefore,
\[ \frac{1 - \epsilon}{1 + \epsilon} \cdot q(f_0 | g_0) \leq p(f_0 | g_0) \leq \frac{1 + \epsilon}{1 - \epsilon} \cdot q(f_0 | g_0). \]

When \( 0 \leq \epsilon \leq \frac{1}{4} \), we have \( (1 + \epsilon)/(1 - \epsilon) \leq 1 + 3\epsilon \) and \( (1 - \epsilon)/(1 + \epsilon) \geq 1 - 3\epsilon \geq 0 \) by straightforward algebra. Hence, for any \( f_0 \in \{0,1\} \) and \( g_0 \in \mathcal{G} \),
\[ (1 - 3\epsilon) q(f_0 | g_0) \leq p(f_0 | g_0) \leq (1 + 3\epsilon) q(f_0 | g_0), \]
by which \( |p(f_0 | g_0) - q(f_0 | g_0)| \leq 3\epsilon \cdot q(f_0 | g_0) \). Thus, for any \( g_0 \in \mathcal{G} \), since \( \epsilon < \frac{1}{6} \) by assumption,
\[ d(g_0) = \sum_{f_0} |p(f_0 | g_0) - q(f_0 | g_0)| \leq 3\epsilon \sum_{f_0} q(f_0 | g_0) = 3\epsilon < \frac{1}{2}. \]

Since \( F \) and \( \tilde{F} \) are binary, we conclude from [20, Theorem 17.3.3] that for any \( g_0 \in \mathcal{G} \),
\[ |H(F|G = g_0) - H(\tilde{F}|G = g_0)| \leq -d(g_0) \log \frac{d(g_0)}{2} \leq -3\epsilon \log \frac{3\epsilon}{2}. \] (129)

Inequality (a) is true because \( x \mapsto -x \log \frac{x}{2} \) is increasing for \( 0 \leq x < \frac{3}{2} \), and \( 0 \leq d(g_0) \leq 3\epsilon < \frac{1}{4} < \frac{3}{2} \) by assumption.

Let \( \Sigma^+ \) denote summation over all \( g_0 \in \mathcal{G} \) for which \( p(g_0) \geq q(g_0) \), and \( \Sigma^- \) denote summation over all \( g_0 \in \mathcal{G} \) for which \( p(g_0) < q(g_0) \). Since \( \Sigma_{g_0} p(g_0) = \Sigma_{g_0} q(g_0) = 1 \), we have
\[ \sum_{g_0}^+ (p(g_0) - q(g_0)) = \sum_{g_0}^- (p(g_0) - q(g_0)) = \frac{1}{2} \sum_{g_0} |p(g_0) - q(g_0)| \leq \frac{\epsilon}{2} \sum_{g_0} q(g_0) = \frac{\epsilon}{2}, \]
where the inequality is by (128). Hence, for any nonnegative function \( h : \mathcal{G} \rightarrow \mathbb{R}^+ \),
\[ \sum_{g_0} (p(g_0) - q(g_0)) h(g_0) = \sum_{g_0}^+ (p(g_0) - q(g_0)) h(g_0) - \sum_{g_0}^- (p(g_0) - q(g_0)) h(g_0) \leq \left( \sup_{g_0} h(g_0) - \inf_{g_0} h(g_0) \right) \cdot \frac{\epsilon}{2} \leq \left( \sup_{g_0} h(g_0) - \inf_{g_0} h(g_0) \right) \cdot \frac{\epsilon}{2}. \] (130)

Therefore,
\[ H(F|G) - H(\tilde{F}|G) \]
\[ = \sum_{g_0} p(g_0) H(F|G = g_0) - \sum_{g_0} q(g_0) H(\tilde{F}|G = g_0) \]
\[ \leq \sum_{g_0} p(g_0) \left( H(\tilde{F}|G = g_0) - 3\epsilon \log \frac{3\epsilon}{2} \right) \]
\[ - \sum_{g_0} q(g_0) H(\tilde{F}|G = g_0) \]
\[ = -3\epsilon \log \frac{3\epsilon}{2} + \sum_{g_0} (p(g_0) - q(g_0)) H(\tilde{F}|G = g_0) \]
\[ \leq -3\epsilon \log \frac{3\epsilon}{2} + \left( \max_{g_0} H(\tilde{F}|G = g_0) - \min_{g_0} H(\tilde{F}|G = g_0) \right) \cdot \frac{\epsilon}{2} \]
\[ \leq \frac{\epsilon}{2} - 3\epsilon \log \frac{3\epsilon}{2}, \]
where (a) is by (129), (b) is by (130), and (c) is because the entropy of a binary random variable assumes values between 0 and 1.

Similarly,
\[ \sum_{g_0} (p(g_0) - q(g_0)) \tilde{h}(g_0) \geq -\left( \sup_{g_0} h(g_0) - \inf_{g_0} h(g_0) \right) \cdot \frac{\epsilon}{2}, \]
Therefore, for any \( x \) as desired.

Thus, we have shown that

\[
\left| H(F|G) - H(\tilde{F}|\tilde{G}) \right| \leq \frac{\varepsilon}{2} - 3\varepsilon \log \frac{3\varepsilon}{2}.
\]

By Lemma 46 below and some algebra, we obtain that

\[
\frac{\varepsilon}{2} - 3\varepsilon \log \frac{3\varepsilon}{2} \leq 2\sqrt{\varepsilon} - \sqrt{2} \varepsilon < \sqrt{8\varepsilon}.
\]

which completes the proof.

**Lemma 46.** For any \( y > 0 \), we have

\[
y(2 - \ln y) \leq 2\sqrt{y}.
\]

**Proof:** This inequality is illustrated in Figure 13. A formal proof follows. The Fenchel dual of \( f(x) = e^x \) [51, p. 105] is

\[
f^*(y) = \sup_{x} (xy - e^x) = \begin{cases} y \ln y - y, & y > 0, \\ 0, & y = 0, \\ \infty, & \text{otherwise}. \end{cases}
\]

Therefore, for any \( x \in \mathbb{R} \) and \( y > 0 \) we have \( xy - e^x \leq y \ln y - y \). Now, set \( x = \frac{1}{2} \ln y \) and rearrange to yield \( y(2 - \ln y) \leq 2\sqrt{y} \) as desired.

**Appendix E**

**Equivalence of the Deterministic and Probabilistic Formulations of Hidden Markov Models**

Recall that in a FAIM process, the observations are a probabilistic function of the state, see (32). However, in Section VIII, we defined the observations of a hidden Markov model as a deterministic function of the state. Seemingly, the deterministic model is less general than the probabilistic FAIM model. As in [18] and [19], we now show that the deterministic and probabilistic models are equivalent.

Using the notation of Section VIII, a hidden Markov model consists of a Markov state \( A_n \) and an observation \( B_n \). In the deterministic model, \( B_n = f(A_n) \), where \( f \) is a deterministic function. In the probabilistic model, there exists a distribution \( q \) such that

\[
P \left( B_n = b \mid A_n = j, B_1^{n-1}, A_1^{t-1} \right) = P \left( B_n = b \mid A_n = j \right) = q(b|j).
\]

One direction of the equivalence is easy: any deterministic model can be thought of a probabilistic model with \( q(|j) \) assuming only the values 0 and 1. To cast the probabilistic model as a deterministic one, observe that by the Markov property and (131), we have

\[
P \left( B_n = b, A_n = j \mid A_{n-1} = i, A_1^{t-2}, B_1^{t-1} \right) = P \left( B_n = b, A_n = j \mid A_{n-1} = i \right) = P \left( A_n = j \mid A_{n-1} = i \right) \cdot P \left( B_n = b \mid A_n = j \right) = p(j|i)q(b|j).
\]

We call a pair \((j, b), j \in A, b \in B, \text{viable if} q(b|j) > 0\). Define a new Markov chain \( C_n \) with states \((j, b)\) whenever \((j, b)\) is a viable pair, and whose transition probability function for any two states \((j, b)\) and \((i, k)\) is \( P \left( C_n = (j, b) \mid C_{n-1} = (i, k) \right) = p(j|i)q(b|j) \). Set \( f : A \times B \rightarrow B \) as the deterministic function that outputs its second argument. That is, \( f(a, b) = b \). This model is deterministic, and is equivalent to the probabilistic one.

We are now almost done; all that remains is to show that \( C_n \) is regular (aperiodic and irreducible) if and only if \( A_n \) is.

**Lemma 47.** Let \( A_n \) be a finite-state homogeneous Markov chain and let \( B_n \) be a probabilistic observation of \( A_n \), as in (131). Then, \( A_n \) is aperiodic and irreducible if and only if \( C_n = (A_n, B_n) \) as defined above is aperiodic and irreducible.

**Proof:** Recall that a finite-state homogeneous Markov chain is aperiodic and irreducible if and only if its transition matrix is primitive. That is, if and only if there exists an integer \( m \) such that the \( m \)-step transition probability from state \( i \) to state \( j \) is positive for any \( i, j \) [34, Theorem 1.4 and Section 4.2], also [32, Section 4.1].

Assume first that \( A_n \) is aperiodic and irreducible. Hence, there exists \( m \) such that \( P \left( A_n = j \mid A_{n-m} = i \right) > 0 \) for all \( i, j \) and \( n \). Therefore, for any viable pairs \((j, b)\) and \((i, k)\),

\[
P \left( C_n = (j, b) \mid C_{n-m} = (i, k) \right) = q(b|j)P \left( A_n = j \mid A_{n-m} = i \right) > 0.
\]

Since the states of \( C_n \) consist only of viable pairs, we conclude that \( C_n \) is aperiodic and irreducible.

Next, assume that \( C_n \) is aperiodic and irreducible. Then, there exists \( m \) such that \( P \left( C_n = (j, b) \mid C_{n-m} = (i, k) \right) > 0 \) for any two viable pairs \((j, b)\) and \((i, k)\), and all \( n \). Therefore, for any \( k \) such that \((i, k)\) is viable (at least one such \( k \) must exist),

\[
P \left( A_n = j \mid A_{n-m} = i \right) = \sum_b P \left( C_n = (j, b) \mid C_{n-m} = (i, k) \right) > 0.
\]

Hence, \( A_n \) is aperiodic and irreducible.

**Example 8.** The Gilbert-Elliott channel [52] is a classic example of a channel with memory. It is defined as follows. The channel may be at one of two states, good and bad. In the good state, the channel is a binary symmetric channel (BSC) with crossover probability \( p \) and in the bad state, the channel is a BSC with crossover probability \( \beta \). The probability of transitioning from the good state to the bad state is \( p \), and the probability of transitioning from the bad state to the good state is \( q \).

\[16\] States for which \( q(b|j) = 0 \) can never appear with positive probability and are therefore removed.
Assuming a symmetric channel input, we construct a deterministic model \( C_n = (S_n, X_n, Y_n) \) with states

\[
\begin{align*}
1 &= (\text{good}, 0, 0), \\
2 &= (\text{good}, 0, 1), \\
3 &= (\text{good}, 1, 0), \\
4 &= (\text{good}, 1, 1), \\
5 &= (\text{bad}, 0, 0), \\
6 &= (\text{bad}, 0, 1), \\
7 &= (\text{bad}, 1, 0), \\
8 &= (\text{bad}, 1, 1).
\end{align*}
\]

The transition probability matrix of \( C_n \) is

\[
M = \begin{bmatrix}
\bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} \\
\bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} \\
\bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} & \bar{p} \\
\bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} \\
\bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} \\
\bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} \\
\bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} & \bar{q} \\
\end{bmatrix}
\]

The possible observations \((X, Y)\) are \((0, 0), (0, 1), (1, 0), \) and \((1, 1)\). The matrices \(M(b), b \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}\) are obtained from \(M\) by replacing all but two columns of \(M\) with zeros. Namely, in \(M(0, 0)\), all but columns 1 and 5 are replaced with zeros; in \(M(0, 1)\) all but columns 2 and 6 are replaced with zeros; in \(M(1, 0)\) all but columns 3 and 7 are replaced with zeros; and in \(M(1, 1)\) all but columns 4 and 8 are replaced with zeros.

\textbf{ACKNOWLEDGMENT}

The authors are grateful to Prof. Rami Atar and Dr. Lele Wang for helpful discussions at the preliminary stages of this work.

\textbf{REFERENCES}


