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 Arıkan'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 Arıkan'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 Arıkan's polar codes that achieve universality over a set of BMS channels. Their first construction combines Arıkan'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 Arıkan 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 Arıkan'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 Arıkan transforms as building blocks. The fast stage consists of multiple applications of Arıkan transforms as in the seminal paper [2]. The slow stage uses Arıkan 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, this universal construction is functionally equivalent to the one presented in [8]. Our goal, however, is to use it over a set of processes with memory.

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 Şaşoğlu'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.

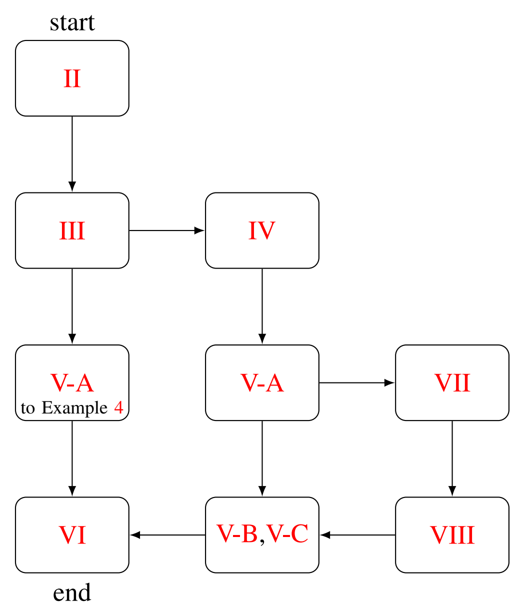

Fig. 1. Roadmap of the various ways to read this paper. All paths start at Section II and end at Section VI.

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 caligraphic letter, e.g., $\mathcal{A}$. The number of elements in a discrete set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. We denote $y_j^k = \begin{bmatrix} y_j & y_{j+1} & \cdots & y_k \end{bmatrix}$ 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 ith 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 ith 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 $\mathbf{a} \overset{f}{\equiv} \mathbf{b}$ 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 $\mathbf{a} \equiv \mathbf{b}$. 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., M. The i, j element of a matrix M is denoted by $(\mathsf{M})_{i,j}$. The ith row of M is denoted by $(\mathsf{M})_{i,:}$ and the jth column of M is denoted by $(\mathsf{M})_{:,j}$. The identity matrix is denoted by $\mathbb{I}$. For matrix M, we denote its set of nonzero rows[1] by $\mathcal{N}_{\mathrm{r}}(\mathsf{M})$ and its set of nonzero columns by $\mathcal{N}_{\mathrm{c}}(\mathsf{M})$. The support $\sigma(\mathsf{M})$ of a matrix M is the set of index pairs (i, j) such that $i \in \mathcal{N}_{\mathrm{r}}(\mathsf{M})$ and $j \in \mathcal{N}_{\mathrm{c}}(\mathsf{M})$.

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 of random variables and $\mathbf{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_m \end{bmatrix}$ is a vector of indices, then $X_{\mathbf{b}} = (X_{b_1}, X_{b_2}, \ldots, X_{b_m})$.

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

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

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x),$$
$$H(X|Y) = -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_{X,Y}(x, y) \log P_{X|Y}(x|y),$$

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

$$h_2(x) = -x \log x - (1 - x) \log(1 - x). \tag{1}$$

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 $X \multimap (Y, Q) \multimap W$ whenever X and W are independent given Y and Q. We then have the following variation of the data processing inequality:

$$X \multimap (Y, Q) \multimap W \Rightarrow I(X; Y|Q) \geq I(X; W|Q). \tag{2}$$

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 $X \rightarrowtail Y$ 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 $X \rightarrowtail Y$ is $H(X|Y)$.

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 dependent observation Y distributed according to $P_{Y|X}(y|x)$.

Definition 2 (s/o-process)**.** A sequence of s/o-pairs $X_i \rightarrowtail Y_i$, $i = 1, 2, \ldots$ is called a *symbol-observation process*, or *s/o-process* in short. We use the notation $X_\star \rightarrowtail Y_\star$.

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 $X_1^N \rightarrowtail Y_1^N$. An s/o-block has a natural indexing: $X_j \rightarrowtail Y_j$ is s/o-pair j of s/o-block $X_1^N \rightarrowtail Y_1^N$. The joint distribution of an s/o-block is given by $P_{X_1^N,Y_1^N}(x_1^N, y_1^N) = P_{X_1^N}(x_1^N) P_{Y_1^N|X_1^N}(y_1^N|x_1^N)$.

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 $X_1^N \rightarrowtail Y_1^N$, the s/o-pairs $X_i \rightarrowtail Y_i$ are identically distributed for all i.

The *conditional entropy rate* of a stationary s/o-process $X_\star \rightarrowtail Y_\star$ is

$$\begin{aligned}\mathcal{H}(X_\star|Y_\star) &\triangleq \lim_{N\to\infty} \frac{1}{N} H(X_1^N|Y_1^N) \\ &= \lim_{N\to\infty} \frac{1}{N} H(X_1^N, Y_1^N) - \lim_{N\to\infty} \frac{1}{N} H(X_1^N).\end{aligned}$$

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 over an alphabet of prime size is possible using the techniques of [6, Chapter 3]. This entails replacing modulo-2 addition with modulo-$|\mathcal{X}|$ addition, where $|\mathcal{X}|$ 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 $X_1^N \rightarrowtail Y_1^N$ is transformed to s/o-block $F_1^N \rightarrowtail G_1^N$ by H-transform f, we mean that:

1) $F_1^N = f(X_1^N)$;
2) $G_i = (F_1^{i-1}, Y_1^N)$, for any i.

Example 1. Arıkan's polar codes [2] are based on H-transforms. In this case, the mapping f is given by $F_1^N = f(X_1^N) = \mathsf{B}_N \mathsf{G}_2^{\otimes n} X_1^N$, where $N = 2^n$, B_N is the $N \times N$ bit-reversal matrix, $\mathsf{G}_2 = \left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$, and $\otimes$ denotes a Kronecker product.

The name "H-transform" is motivated by the equality

$$H(X_1^N|Y_1^N) = H(F_1^N|Y_1^N) = \sum_{i=1}^{N} H(F_i|G_i). \tag{3}$$

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 $X_1^N \rightarrowtail Y_1^N$, with H-transform $F_1^N \rightarrowtail G_1^N$. We wish to recover the symbols X_1^N from the observations Y_1^N. We denote the recovered symbols with a hat, $(\hat{\cdot})$. That is, $\hat{X}_1^N = \Phi(Y_1^N)$, where $\Phi(\cdot)$ is the algorithm for recovery. We assess Φ by its error probability, $\mathbb{P}(\hat{X}_1^N \neq X_1^N)$. H-transforms, thanks to (3), naturally give rise to a sequential algorithm called *successive cancellation*.

Rather than computing $\hat{X}_1^N$ from Y_1^N directly, we may compute $\hat{F}_1^N$ from Y_1^N. By the properties of the H-transform, there exists a mapping f, with inverse f^{-1}, such that $X_1^N = f^{-1}(F_1^N)$. Any algorithm for recovering F_1^N from Y_1^N is equivalent to an algorithm for recovering X_1^N from Y_1^N. For, if $\hat{F}_1^N = \Phi(Y_1^N)$ we can define $\hat{X}_1^N = f^{-1}(\hat{F}_1^N) = f^{-1}(\Phi(Y_1^N))$ and vice versa. Since $\mathbb{P}(\hat{F}_1^N \neq F_1^N) = \mathbb{P}(\hat{X}_1^N \neq X_1^N)$, we concentrate on an algorithm to recover F_1^N.

One approach is to compute $\hat{F}_1^N$ sequentially as follows. Let Φ_i be a maximum-likelihood decoder of F_i from G_i. Compute $\hat{F}_1 = \Phi_1(\hat{G}_1)$, where $\hat{G}_1 = G_1 = Y_1^N$; then, assuming that $\hat{F}_1 = F_1$, form $\hat{G}_2 = (\hat{F}_1, Y_1^N)$ 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 $\hat{F}_1^N$ sequentially from $\hat{G}_1^N$ using successive cancellation as described above is the same as if a genie had replaced $\hat{G}_i$ with G_i at every step. That is,

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

(To see this, observe that if $\Phi_i(G_i) = F_i$ for all $i < i_0$ and $\Phi_{i_0}(G_{i_0}) \neq F_{i_0}$ 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 $\mathcal{L}, \mathcal{H} \subseteq \{1, 2, \ldots, N\}$ be two index sets. An H-transform f is $(\eta, \mathcal{L}, \mathcal{H})$-*monopolarizing* for a family of s/o-processes if for any s/o-block $X_1^N \rightarrowtail Y_1^N$ 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 $i \in \mathcal{H}$, where s/o-block $F_1^N \rightarrowtail G_1^N$ 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

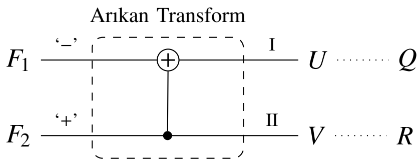

Fig. 2. Illustration of an Arıkan transform. It transforms two input symbols, U (input-I) and V (input-II) to two output symbols, F_1 (output '−') and F_2 (output '+').

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 φ_{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: $U_1^{N_n} = f_n(X_1^{N_n})$ and $V_1^{N_n} = f_n(X_{N_n+1}^{2N_n})$. Then, form $F_1^{N_{n+1}} = \varphi_{n+1}(U_1^{N_n}, V_1^{N_n}) = f_{n+1}(X_1^{N_{n+1}})$.

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:

$$\begin{cases} \mathrm{I}: U \rightarrowtail Q \\ \mathrm{II}: V \rightarrowtail R \end{cases} \Rightarrow \begin{cases} \text{'}-\text{'}: \underbrace{U+V}_{F_1} \rightarrowtail \underbrace{(Q,R)}_{G_1} \\ \text{'}+\text{'}: \underbrace{V}_{F_2} \rightarrowtail \underbrace{(F_1,Q,R)}_{G_2} \end{cases}.$$

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

For an Arıkan transform, we obtain

$$\begin{aligned} H(F_1|G_1) + H(F_2|G_2) &= H(F_1^2|Q,R) \\ &= H(U,V|Q,R) \le H(U|Q) + H(V|R). \end{aligned}$$

The inequality is because the s/o-pairs $U \rightarrowtail Q$ and $V \rightarrowtail R$ are generally dependent. Informally, Arıkan transforms facilitate polarization if one can show that $H(F_1|G_1) \ge \max\{H(U|Q), H(V|R)\}$ 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 startegy.

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, $n \ge 0$. See Figure 3 for an illustration.

$$\begin{aligned} [\mathrm{lat}_1(n)] &\triangleq \{i \mid \quad 1 \le i \le L_n\}, & (4a)\\ [\mathrm{lat}_2(n)] &\triangleq \{i \mid L_n + M_n + 1 \le i \le N_n\}, & (4b)\\ [\mathrm{lat}(n)] &\triangleq [\mathrm{lat}_1(n)] \cup [\mathrm{lat}_2(n)], & (4c)\\ [\mathrm{med}_-(n)] &\triangleq \{i \mid i = L_n + 2k - 1,\ 1 \le k \le M_n/2\}, & (4d)\\ [\mathrm{med}_+(n)] &\triangleq \{i \mid i = L_n + 2k, \quad 1 \le k \le M_n/2\}, & (4e)\\ [\mathrm{med}(n)] &\triangleq [\mathrm{med}_-(n)] \cup [\mathrm{med}_+(n)]. & (4f) \end{aligned}$$

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

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

- $i \in [\mathrm{lat}(n)] \quad \Rightarrow$ symbol i is lateral;
- $i \in [\mathrm{med}(n)] \quad \Rightarrow$ symbol i 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.[2]

- 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 $X_1^{N_0} \rightarrowtail Y_1^{N_0}$ into an s/o-block $F_1^{N_0} \rightarrowtail G_1^{N_0}$, where, for $1 \le i \le N_0$,

$$\begin{aligned} F_i &= X_i, & (5a)\\ G_i &= (F_1^{i-1}, Y_1^{N_0}). & (5b) \end{aligned}$$

We 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 φ_{n+1} from two length-N_n vectors to a single length-$N_{n+1} = 2N_n$ vector. The mapping φ_{n+1} is defined in (9) and (10) below.

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

$$\begin{aligned} U_1^{N_n} &= f_n(X_1^{N_n}), \quad & Q_i &= (U_1^{i-1}, Y_1^{N_n}), \quad & 1 \le i \le N_n, \quad (6a)\\ V_1^{N_n} &= f_n(X_{N_n+1}^{2N_n}), \quad & R_i &= (V_1^{i-1}, Y_{N_n+1}^{2N_n}), \quad & 1 \le i \le N_n. \quad (6b) \end{aligned}$$

[2]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 $n = 0$.

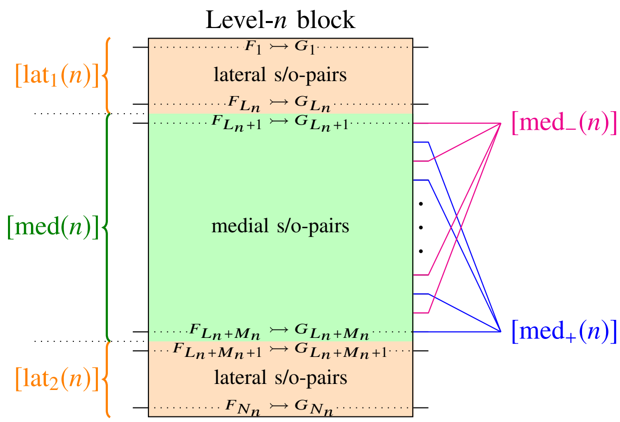

Fig. 3. Index sets in level n of the basic slow transform. A Level-n block comprises $N_n = 2L_n + M_n$ s/o-pairs. The first L_n and the last L_n s/o-pairs are lateral s/o-pairs and the remaining M_n s/o-pairs are medial s/o-pairs.

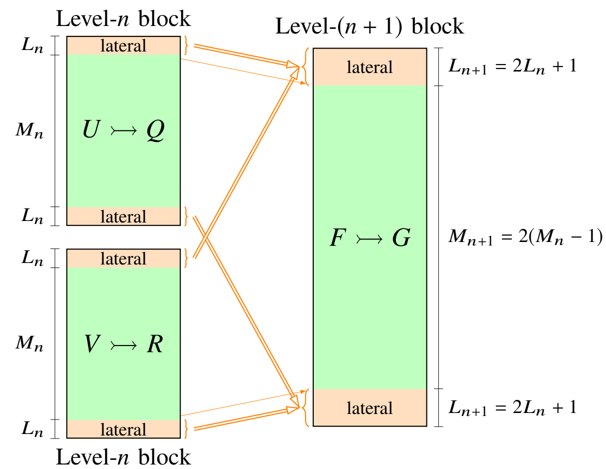

Fig. 4. A schematic description of forming lateral s/o-pairs of a level-$(n+1)$ block from two level-n blocks.

Denoting $N_{n+1} = 2N_n$, we obtain the level-$(n+1)$ transformed s/o-block

$$F_1^{N_{n+1}} = \varphi_{n+1}(U_1^{N_n}, V_1^{N_n}) = f_{n+1}(X_1^{N_{n+1}}), \tag{7a}$$

$$G_i = (F_1^{i-1}, Y_1^{N_{n+1}}), \quad 1 \le i \le N_{n+1}. \tag{7b}$$

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

$$L_{n+1} = 2L_n + 1 \tag{8a}$$

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

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 $n \ge 1$. Therefore, for any $n \ge 1$, the set $[\mathrm{med}_-(n)]$ is the set of even indices of $[\mathrm{med}(n)]$ and the set $[\mathrm{med}_+(n)]$ is the set of odd indices of $[\mathrm{med}(n)]$.

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

$$i \in [\mathrm{lat}(n+1)] \Rightarrow F_i = \begin{cases} U_j, & i = 2j-1, \\ V_j, & i = 2j. \end{cases} \tag{9}$$

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

$$F_{L_{n+1}+M_{n+1}+1} = F_{2(L_n+M_n)} = V_{L_n+M_n}.$$

The medial symbols of a level-$(n+1)$ block are formed using Arıkan transforms, as illustrated in Figure 5. That is, medial symbols of a level-$(n+1)$ block are computed according to:

$$i \in [\mathrm{med}(n+1)] \Rightarrow$$

$$F_i = \begin{cases} U_{j+1} + V_j, & i = 2j, \\ V_j, & i = 2j+1,\ j \in [\mathrm{med}_-(n)], \\ U_{j+1}, & i = 2j+1,\ j \in [\mathrm{med}_+(n)]. \end{cases} \tag{10}$$

We emphasize that

$$i \in [\mathrm{med}(n+1)] \Leftrightarrow \left\{ \left\lfloor \frac{i}{2} \right\rfloor, \left\lfloor \frac{i}{2} \right\rfloor + 1 \right\} \in [\mathrm{med}(n)]. \tag{11}$$

That is, medial symbols of a level-$(n+1)$ BST are generated by combining medial symbols of level-n BSTs. This can be seen either from Figure 5 or from (4), (8), and (10). In particular, (10) and (11) imply that for any $n \ge 0$,

$$i \in [\mathrm{med}_-(n+1)] \Leftrightarrow i = 2j, \qquad j \in [\mathrm{med}(n)],\ j \ne N_n - L_n,$$

$$i \in [\mathrm{med}_+(n+1)] \Leftrightarrow i = 2j+1,\ j \in [\mathrm{med}(n)],\ j \ne N_n - L_n.$$

Figure 5 makes it clear that the medial symbols of a level-$(n+1)$ block are formed in pairs. Overall, $M_n - 1$ Arıkan transforms are performed in forming the medial symbols of a level-$(n+1)$ block. Recall that an Arıkan transform has two inputs, I and II, see Figure 2. In each Arıkan transform, input-I is a symbol from $[\mathrm{med}_+(n)]$ of one level-n block and input-II is a symbol from $[\mathrm{med}_-(n)]$ of the other level-n block. The blocks *alternate* between successive Arıkan transforms: look at $F_{2L_n+2}, F_{2L_n+3}, F_{2L_n+4}$, and F_{2L_n+5} in Figure 5.

We saw above that the first medial symbol of the first level-n block and the last medial symbol of the second level-n block become lateral symbols of the level-$(n+1)$ block; they do not participate in forming medial symbols of the level-$(n+1)$ block. This explains why the index of U leads by one the index of V in (10).

When $2j \in [\mathrm{med}(n+1)]$, F_{2j} and F_{2j+1} are the outputs of an Arıkan transform of U_{j+1} and V_j. The expression for F_{2j} is always the same: $F_{2j} = U_{j+1} + V_j$. The expression for F_{2j+1} depends on which of U_{j+1} or V_j is input-II of the Arıkan transform. One of j and $j+1$ is in $[\mathrm{med}_-(n)]$ and the other is in $[\mathrm{med}_+(n)]$. Since we form medial symbols using Arıkan transforms with input-II symbols from $[\mathrm{med}_-(n)]$ of a level-n block, F_{2j+1} is assigned according to the classification of j. Observe that for any $n \ge 1$, by Remark 1, the condition "$j \in [\mathrm{med}_-(n)]$" is the same as "$j$ is even", and the condition "$j \in [\mathrm{med}_+(n)]$" is the same as "$j$ is odd."

We pause momentarily to introduce some terminology that will be useful in the sequel.

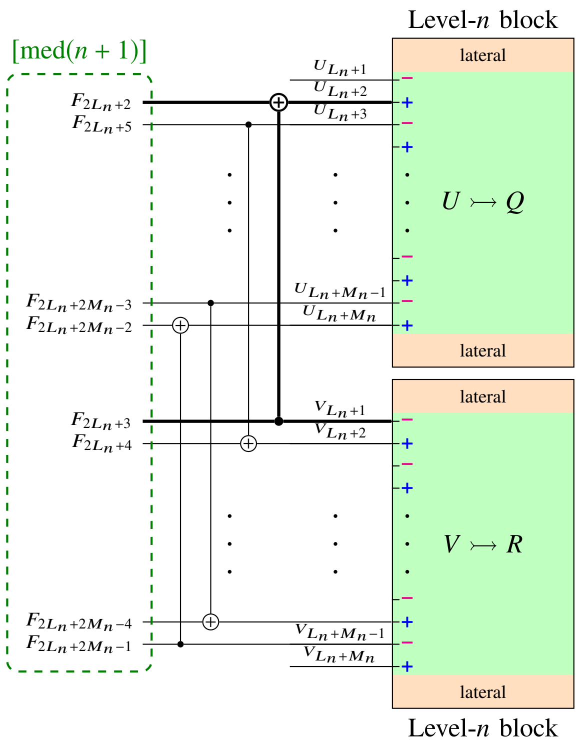

Fig. 5. Forming the medial symbols of level $n+1$ of the basic slow transform. Arıkan transforms are used with a symbol from $[\mathrm{med}_+(n)]$ of one block as their input-I and a symbol from $[\mathrm{med}_-(n)]$ of the other block as their input-II. One Arıkan transform is highlighted using thicker edges.

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 $i = 2j \in [\mathrm{med}(n+1)]$. Then, $i+1 \in [\mathrm{med}(n+1)]$ as well, see (4) and Remark 1. Medial symbols F_i 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 $F_1^{N_{n+1}}$. We define the function φ_{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 φ_{n+1} is defined according to (9) and (10). Then, for any $n \geq 0$,*

$$i \in [\mathrm{lat}(n+1)] \Rightarrow G_i \equiv \begin{cases} (Q_j, R_j), & i = 2j-1, \\ (Q_{j+1}, R_j), & i = 2j \neq 2N_n, \\ (F_{i-1}, Q_{N_n}, R_{N_n}), & i = 2N_n \end{cases} \tag{12}$$

and

$$i \in [\mathrm{med}(n+1)] \Rightarrow G_i \equiv \begin{cases} (Q_{j+1}, R_j), & i = 2j, \\ (F_{i-1}, Q_{j+1}, R_j), & i = 2j+1. \end{cases} \tag{13}$$

Proof: By construction, for $1 \leq j \leq N_n$, we have

$$Q_j = (U_1^{j-1}, Y_1^{N_n}), \quad R_j = (V_1^{j-1}, Y_{N_n+1}^{2N_n}).$$

Since

$$G_i = (F_1^{i-1}, Y_1^{2N_n}),$$

we need only show that there is a one-to-one mapping between the non-Y portions of the right-hand-sides of (12) and (13) to F_1^{i-1}. We proceed in cases, based on the index i in the level-$(n+1)$ block.

Case 1: $i \in [\mathrm{lat}_1(n+1)]$ — the first half of the lateral set, see (4a).
In this case, to show (12) it suffices to establish

$$F_1^{i-1} \equiv \begin{cases} (U_1^{j-1}, V_1^{j-1}), & i = 2j-1, \\ (U_1^{j}, V_1^{j-1}), & i = 2j. \end{cases} \tag{14}$$

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 $i \in [\mathrm{lat}_1(n+1)]$.

Case 2: $i \in [\mathrm{med}(n+1)]$ — the medial set, see (4f).
In this case, to show (13) it suffices to establish

$$F_1^{i-1} \equiv \begin{cases} (U_1^{j}, V_1^{j-1}), & i = 2j, \\ (F_{i-1}, U_1^{j}, V_1^{j-1}), & i = 2j+1. \end{cases} \tag{15}$$

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), $F_1^{i-1} \equiv (U_1^{L_n+1}, V_1^{L_n})$, and trivially $F_1^{i} \equiv (F_i, U_1^{L_n+1}, V_1^{L_n})$. This implies (15) for the first two medial indices. We continue by induction. Assume that for $i = 2j \in [\mathrm{med}(n+1)]$ we have $F_1^{2j-1} \equiv (U_1^{j}, V_1^{j-1})$. Trivially, $F_1^{2j} \equiv (F_{2j}, U_1^{j}, V_1^{j-1})$; hence (15) holds for $i+1$ as well. By (10),

$$\begin{aligned} F_1^{2(j+1)-1} &\equiv (F_1^{2j-1}, F_{2j}, F_{2j+1}) \\ &\equiv (F_1^{2j-1}, U_{j+1}, V_j) \\ &\equiv (U_1^{j+1}, V_1^{j}), \end{aligned} \tag{16}$$

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

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

Case 3: $i \in [\mathrm{lat}_2(n+1)]$ — the second half of the lateral set, see (4b).
In this case, to show (12) it suffices to establish

$$F_1^{i-1} \equiv \begin{cases} (U_1^{j-1}, V_1^{j-1}), & i = 2j-1, \\ (U_1^{j}, V_1^{j-1}), & i = 2j \neq 2N_n, \\ (F_{i-1}, U_1^{N_n-1}, V_1^{N_n-1}), & i = 2N_n. \end{cases} \tag{17}$$

If i is the first lateral index in $[\mathrm{lat}_2(n+1)]$, 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, $F_1^{2(L_n+M_n)-1} \equiv (U_1^{L_n+M_n}, V_1^{L_n+M_n-1})$. For any other index $i \in [\mathrm{lat}_2(n+1)]$, 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

$$\alpha_n \triangleq \frac{M_n}{2L_n + M_n}. \tag{18}$$

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

$$M_0 \geq \left\lceil \frac{2(1+\alpha L_0)}{1-\alpha} \right\rceil,$$

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 $n \geq 0$,

$$\begin{aligned} L_n &= 2^n(L_0 + (1 - 2^{-n})) \\ M_n &= 2^n(M_0 - 2(1 - 2^{-n})). \end{aligned}$$

Therefore, recalling that $N_0 = 2L_0 + M_0$,

$$\alpha_n = \frac{M_n}{2L_n + M_n} = \frac{M_0 - 2(1-2^{-n})}{2L_0 + M_0} = \alpha_0 - \frac{2(1-2^{-n})}{N_0}.$$

This implies that

$$\alpha_n \geq \alpha_0 - \frac{2}{N_0} = \frac{M_0 - 2}{M_0 + 2L_0}.$$

The right-hand side is an increasing function of M_0, since its derivative with respect to M_0 is $2(1+L_0)/(2L_0+M_0)^2 > 0$. It remains to find m_0 such that $(m_0-2)/(m_0+2L_0) = \alpha$. Then, for any $M_0 \geq \lceil m_0 \rceil$, 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 Şaşoğlu-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 η and family of s/o-processes, the BST is $(\eta, \mathcal{L}, \mathcal{H})$-monopolarizing, with $\mathcal{L} = [\mathrm{med}_+(n)]$ and $\mathcal{H} = [\mathrm{med}_-(n)]$, 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 $[\mathrm{med}_-(n)]$ have approximately the same conditional entropy; the same is true for all symbols in $[\mathrm{med}_+(n)]$. If n is sufficiently large, one set will have polarized (e.g., the conditional entropies of s/o-pairs in $[\mathrm{med}_-(n)]$ 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 $(\eta, \mathcal{L}, \mathcal{H})$-monopolarizing 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 η; 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 $(\eta, \mathcal{L}, \mathcal{H})$-monopolarizing[3] 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 η. 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 $\hat{N} = 2^{\hat{n}}$ 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 either $\mathcal{L}$ 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 jth Arıkan transform operates only on the jth 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.

[3] If 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.

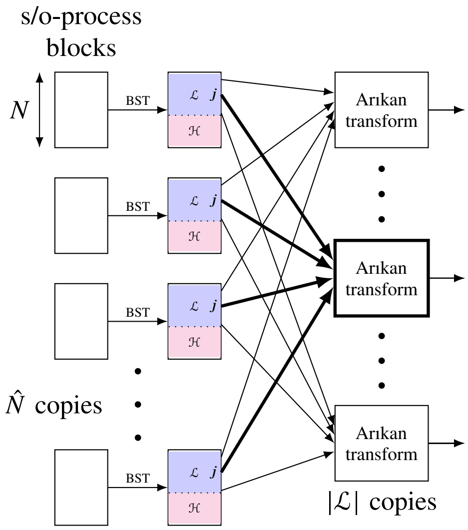

Fig. 6. Illustration of the fast stage for a channel-coding application. First, $\hat{N}$ length-N blocks of the s/o-process are transformed using BSTs of length N. Then, $|\mathcal{L}|$ Arıkan trasforms of length $\hat{N}$ are applied. The jth Arıkan transform (in bold) operates on the jth medial s/o-pair in $\mathcal{L}$ from each length-N BST.

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, $Z_{n+1} \leq \kappa Z_n^{b_i}$, 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].[4] In Appendix A we show that if η 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 η 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 η 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 only on previous transformed symbols; we denote this set by $\mathcal{D}$. In the Honda-Yamamoto scheme, data is transmitted only over the indices in $\mathcal{D}$. We refer the reader to [21] for details on this scheme. The rate of codes produced this way is $|\mathcal{D}|/N$. Since the input distribution is known and fixed, the set of transformed symbols with very low total variation distance as above is fixed as well and can be determined in advance.

Continuing our channel-coding example, for each s/o-process in the subfamily, the following is true. The Bhattacharyya parameter of almost all s/o-pairs after the fast polarization stage is universally less than $2^{-\hat{N}^{\beta}}$, for a fixed $\beta < 1/2$ and $\hat{N}$ large enough, [2], [13], [14]. In fact, this is true also for the bounds on the Bhattacharyya parameter discussed above, see Appendix A. Since the overall code is of length $N \cdot \hat{N}$, by the union bound, when $\hat{N}$ is large enough, the block error probability of this scheme under successive-cancellation decoding is upper-bounded by $N \cdot \hat{N} \cdot 2^{-\hat{N}^{\beta'}}$. Here, $\beta' < \beta < 1/2$ encompasses any losses in generating the desired input distribution in the Honda-Yamamoto scheme, see [14], [21]. Since n is fixed, so is $N = N(n)$, and thus the error probability vanishes as $\hat{N}$ grows.

In the above discussion, we exemplified the case of universal codes for channel-coding applications. The universal polar code can also be used for source-coding applications, using the set $\mathcal{H}$ instead of $\mathcal{L}$ as the basis for the fast polarization stage.

Remark 2. In the memoryless case there exists a large body of work regarding construction of polar codes, [22]–[30]. The construction of polar codes for processes with memory was not addressed in [13], [14], and remains an open problem. A distinct advantage of the universal polar construction is that it can be made to sidestep the construction process for processes with memory.

[4] Similar relationships for the total variation distance also hold, see [14, Proposition 4 and Proposition 12].

IV. Properties of the BST and a Variation

In this section we explore some of the properties of the BST. We also introduce a variation of the BST, the Observation-truncated BST. We will call upon these when analyzing the BST in Section V-C.

A. Properties of the BST

We now explore some properties of the BST that will be useful in the sequel. To this end, throughout this section we assume that BSTs are initialized with parameters L_0 and M_0. A level-0 BST is thus of length $N_0 = 2L_0 + M_0$, and a level-n BST is of length $N_n = 2^n N_0$.

Since $N_n = 2^n N_0$, we say that a level-n BST is formed from 2^n level-0 BSTs. We call each level-0 BST a b-block, and we number them sequentially. The b-block numbered ℓ contains s/o-pairs with indices $(\ell - 1)N_0 + k$, $1 \leq k \leq N_0$. The following definition names both ℓ and k.

Definition 7 (b-block number and b-index)**.** In a level-n BST, an index j is a number between 1 and N_n. We write it in the form

$$j = (\ell - 1)N_0 + k, \quad 1 \leq \ell \leq 2^n, \quad 1 \leq k \leq N_0. \tag{19}$$

We call ℓ the *b-block number* and k the *b-index* that correspond to index j.[5]

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 $X_1^{N_n}$. 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* $\mathbf{b}$ of a medial index i is a *row* vector whose ℓth entry is the base-ancestor of i from b-block ℓ. Therefore,

$$(\mathbf{b})_\ell = (\ell - 1)N_0 + k \tag{20}$$

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

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

$$(\bar{\mathbf{b}})_\ell = (\mathbf{b})_\ell - (\ell - 1)N_0, \quad 1 \leq \ell \leq 2^n, \tag{21}$$

where n is the level of i. This vector contains in its ℓth entry the b-index of i's base-ancestor in the ℓth 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 medial index.

To motivate the definition of the base-vector, assume momentarily that the s/o-process being transformed were memoryless. If we tried to recover some transformed symbol F_i using successive-cancellation decoding, we could discard all observations except for those whose indices are in the base-vector. That is, in the memoryless case

$$\mathbb{P}\left(F_i = 0 \mid F_1^{i-1}, Y_1^{N_n}\right) = \mathbb{P}\left(F_i = 0 \mid F_1^{i-1}, Y_{\mathbf{b}}\right),$$

where $Y_{\mathbf{b}} = \{Y_{(\mathbf{b})_1}, Y_{(\mathbf{b})_2}, \ldots, Y_{(\mathbf{b})_{2^n}}\}$. 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 $i \in [\mathrm{med}(n)]$, we use $\mathbf{b}_i^{(n)}$ and $\bar{\mathbf{b}}_i^{(n)}$ to denote the base-vector and modulo-base-vector, respectively.

[5] The letter 'b' in these names stands for 'base,' as the BST may be thought of as consisting of 2^n "base blocks" of length N_0.

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:

$$\bar{\mathbf{b}}_i^{(0)} = \begin{bmatrix} i \end{bmatrix}.$$

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

$$\bar{\mathbf{b}}_i^{(n+1)} = \begin{bmatrix} \bar{\mathbf{b}}_{j+1}^{(n)} & \bar{\mathbf{b}}_j^{(n)} \end{bmatrix}, \quad j = \left\lfloor \frac{i}{2} \right\rfloor. \tag{22}$$

Recall from Remark 1 that if $i \in [\mathrm{med}_-(n+1)]$, 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 $N_0 = 2L_0 + M_0 = 12$. 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

$$\bar{\mathbf{b}}_8^{(1)} = \bar{\mathbf{b}}_9^{(1)} = \begin{bmatrix} 5 & 4 \end{bmatrix}, \quad \bar{\mathbf{b}}_{10}^{(1)} = \bar{\mathbf{b}}_{11}^{(1)} = \begin{bmatrix} 6 & 5 \end{bmatrix},$$

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,

$$\bar{\mathbf{b}}_{16}^{(2)} = \bar{\mathbf{b}}_{17}^{(2)} = \begin{bmatrix} 5 & 4 & 5 & 4 \end{bmatrix}, \quad \bar{\mathbf{b}}_{18}^{(2)} = \bar{\mathbf{b}}_{19}^{(2)} = \begin{bmatrix} 6 & 5 & 5 & 4 \end{bmatrix}.$$

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

$$\begin{aligned} \bar{\mathbf{b}}_{32}^{(3)} = \bar{\mathbf{b}}_{33}^{(3)} &= \begin{bmatrix} 5 & 4 & 5 & 4 & 5 & 4 & 5 & 4 \end{bmatrix}, \\ \bar{\mathbf{b}}_{34}^{(3)} = \bar{\mathbf{b}}_{35}^{(3)} &= \begin{bmatrix} 6 & 5 & 5 & 4 & 5 & 4 & 5 & 4 \end{bmatrix}. \end{aligned}$$

Computing a base-vector, say $\mathbf{b}_{35}^{(3)}$, is easily done using (21):

$$\mathbf{b}_{35}^{(3)} = \begin{bmatrix} 6 & 17 & 29 & 40 & 53 & 64 & 77 & 88 \end{bmatrix}.$$

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

Let $n \leq m$. Fix some $i \in [\mathrm{med}(m)]$ 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:

$$\begin{aligned} \bar{\mathbf{b}}_{34}^{(3)} &= \begin{bmatrix} [6 \quad 5] & [5 \quad 4] & [5 \quad 4] & [5 \quad 4] \end{bmatrix} \\ &= \begin{bmatrix} \bar{\mathbf{b}}_{10}^{(1)} & \bar{\mathbf{b}}_9^{(1)} & \bar{\mathbf{b}}_9^{(1)} & \bar{\mathbf{b}}_8^{(1)} \end{bmatrix}. \end{aligned}$$

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 $i, i+1 \in [\mathrm{med}(n)]$ and any $1 \leq \ell \leq 2^n$ we have*

$$(\bar{\mathbf{b}}_{i+1}^{(n)})_\ell \geq (\bar{\mathbf{b}}_i^{(n)})_\ell. \tag{23}$$

Proof: This follows from (22) by straightforward induction. Specifically, note that if the index i on the left-hand-side of

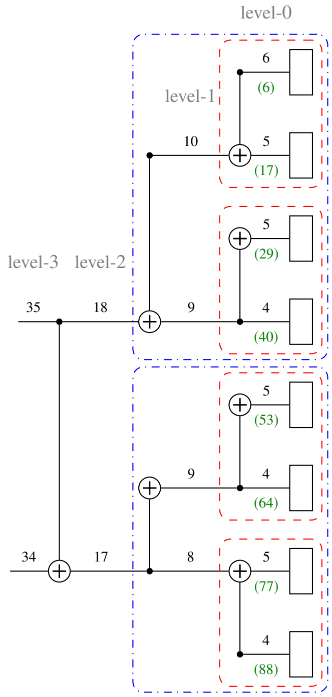

Fig. 7. A portion of a level-3 BST, initialized with $L_0 = 3, M_0 = 6$. The base-vector $\mathbf{b}_{34}^{(3)} = \mathbf{b}_{35}^{(3)}$ is illustrated. The rectangles denote level-0 BSTs. Level-1 BSTs are delimited with dashed lines (in red) and level-2 BSTs are delimited with dash-dotted lines (in blue). Above each line, we show its index with respect to its relevant-level BST (the rightmost are level-0). The level-0 indices are also b-indices; below them we noted in parentheses (in green) their respective indices in a level-3 BST.

(22) increases, the indices j and $j+1$ on the right-hand-side cannot decrease. ■

Lemma 4. *For any $i \in [\mathrm{med}(n)]$ and any $1 \le \ell \le 2^n$ we have*

$$\left\lfloor \frac{i}{2^n} \right\rfloor = (\bar{\mathbf{b}}_i^{(n)})_{2^n} \le (\bar{\mathbf{b}}_i^{(n)})_\ell \le (\bar{\mathbf{b}}_i^{(n)})_1 = 1 + \left\lceil \frac{i-1}{2^n} \right\rceil. \tag{24}$$

In words, for any medial index i, the first element of $\bar{\mathbf{b}}_i^{(n)}$ is its maximal, which equals $1 + \lceil (i-1) \cdot 2^{-n} \rceil$, and the last element of $\bar{\mathbf{b}}_i^{(n)}$ is its minimal, which equals $\lfloor i \cdot 2^{-n} \rfloor$.

Proof: The proof consists of several steps, all proved using induction. First, we prove *claim 1*: $(\bar{\mathbf{b}}_i^{(n)})_1 \ge (\bar{\mathbf{b}}_i^{(n)})_\ell \ge (\bar{\mathbf{b}}_i^{(n)})_{2^n}$ for any $i \in [\mathrm{med}(n)]$ and $1 \le \ell \le 2^n$. Then, we will establish the formulas for computing the values of these elements.

Proof of Claim 1: For $n = 0$ claim 1 is trivially true, as for any $i \in [\mathrm{med}(0)]$, $\bar{\mathbf{b}}_i^{(0)}$ is a singleton. Assume that claim 1 holds for some $n \ge 0$; we will establish that it is true also for $n+1$. Let $i \in [\mathrm{med}(n+1)]$. Then, by (22), $\bar{\mathbf{b}}_i^{(n+1)} = \begin{bmatrix} \bar{\mathbf{b}}_{j+1}^{(n)} & \bar{\mathbf{b}}_j^{(n)} \end{bmatrix}$, where $j = \lfloor i/2 \rfloor$. By the induction hypothesis,

$$\begin{aligned}(\bar{\mathbf{b}}_i^{(n+1)})_1 = (\bar{\mathbf{b}}_{j+1}^{(n)})_1 &\ge (\bar{\mathbf{b}}_{j+1}^{(n)})_\ell \ge (\bar{\mathbf{b}}_{j+1}^{(n)})_{2^n}, \\ (\bar{\mathbf{b}}_j^{(n)})_1 &\ge (\bar{\mathbf{b}}_j^{(n)})_\ell \ge (\bar{\mathbf{b}}_j^{(n)})_{2^n} = (\bar{\mathbf{b}}_i^{(n+1)})_{2^{n+1}}\end{aligned}$$

for any $1 \le \ell \le 2^n$. By Lemma 3, $(\bar{\mathbf{b}}_{j+1}^{(n)})_\ell \ge (\bar{\mathbf{b}}_j^{(n)})_\ell$ for any $1 \le \ell \le 2^n$. Therefore,

$$(\bar{\mathbf{b}}_i^{(n+1)})_1 \ge (\bar{\mathbf{b}}_i^{(n+1)})_\ell \ge (\bar{\mathbf{b}}_i^{(n+1)})_{2^{n+1}}$$

for any $1 \le \ell \le 2^{n+1}$, thereby proving claim 1.

Proof of the right-hand side of (24): For $n = 0$ and any $i \in [\mathrm{med}(0)]$, trivially $(\bar{\mathbf{b}}_i^{(0)})_1 = 1 + \lceil (i-1) \cdot 2^{-0} \rceil = i$. Assume that the right-hand side of (24) holds for some $n \ge 0$; we will show it holds for $n+1$ as well. Let $i \in [\mathrm{med}(n+1)]$; by (22), $(\bar{\mathbf{b}}_i^{(n+1)})_1 = (\bar{\mathbf{b}}_{j+1}^{(n)})_1$, where $j = \lfloor i/2 \rfloor$. Now, observe that for natural i,

$$\left\lfloor \frac{i}{2} \right\rfloor = \left\lceil \frac{i-1}{2} \right\rceil.$$

Therefore,

$$\begin{aligned}(\bar{\mathbf{b}}_i^{(n+1)})_1 = (\bar{\mathbf{b}}_{\lfloor i/2 \rfloor + 1}^{(n)})_1 &\overset{(a)}{=} 1 + \left\lceil \frac{\lfloor i/2 \rfloor}{2^n} \right\rceil = 1 + \left\lceil \frac{\lceil (i-1)/2 \rceil}{2^n} \right\rceil \\ &\overset{(b)}{=} 1 + \left\lceil \frac{(i-1)/2}{2^n} \right\rceil = 1 + \left\lceil \frac{i-1}{2^{n+1}} \right\rceil,\end{aligned}$$

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 $i \in [\mathrm{med}(0)]$, trivially $(\bar{\mathbf{b}}_i^{(0)})_{2^0} = \lfloor i \cdot 2^{-0} \rfloor = i$. Assume that the left-hand side of (24) holds for some $n \ge 0$; we will show it holds for $n+1$ as well. Let $i \in [\mathrm{med}(n+1)]$; by (22), $(\bar{\mathbf{b}}_i^{(n+1)})_{2^{n+1}} = (\bar{\mathbf{b}}_j^{(n)})_{2^n}$, where $j = \lfloor i/2 \rfloor$. Therefore,

$$(\bar{\mathbf{b}}_i^{(n+1)})_{2^{n+1}} = (\bar{\mathbf{b}}_{\lfloor i/2 \rfloor}^{(n)})_{2^n} \overset{(a)}{=} \left\lfloor \frac{\lfloor i/2 \rfloor}{2^n} \right\rfloor \overset{(b)}{=} \left\lfloor \frac{i/2}{2^n} \right\rfloor = \left\lfloor \frac{i}{2^{n+1}} \right\rfloor,$$

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

Corollary 5. *If $n \ge 1$ then for any $i \in [\mathrm{med}(n)]$,*

$$1 \le \max_\ell (\bar{\mathbf{b}}_i^{(n)})_\ell - \min_\ell (\bar{\mathbf{b}}_i^{(n)})_\ell \le 2.$$

Proof: This is an immediate consequence of Lemma 4. Specifically, if $\lfloor i/2^n \rfloor = r$ then

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

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

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 $F_1^{N_n} \rightarrowtail G_1^{N_n}$ be the output of a level-n BST with parameters L_0 and M_0 of s/o-block $X_1^{N_n} \rightarrowtail Y_1^{N_n}$. Recall that $X_i \in \mathcal{X} = \{0,1\}$ and $Y_i \in \mathcal{Y}$ for any i, where $\mathcal{Y}$ is some finite alphabet. For any $i \in [\mathrm{med}(n)]$ there exist functions

$$\begin{aligned}f_{n,i} &: \mathcal{X}^{N_n} \to \mathcal{X}, \\ g_{n,i} &: \mathcal{X}^{N_n} \times \mathcal{Y}^{N_n} \to \mathcal{X}^{i-1} \times \mathcal{Y}^{N_n},\end{aligned}$$

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

$$f_{0,i}(X_1^{N_0}) = X_i, \tag{25a}$$

$$g_{0,i}(X_1^{N_0}, Y_1^{N_0}) = (X_1^{i-1}, Y_1^{N_0}). \tag{25b}$$

Recursion for $f_{n+1,i}$ for any $i \in [\mathrm{med}(n+1)]$:

$$f_{n+1,i}(X_1^{N_{n+1}}) = \begin{cases} f_{n,j+1}(X_1^{N_n}) + f_{n,j}(X_{N_n+1}^{2N_n}), & i = 2j, \\ f_{n,j}(X_{N_n+1}^{2N_n}), & i = 2j+1,\ j \in [\mathrm{med}_-(n)], \\ f_{n,j+1}(X_1^{N_n}), & i = 2j+1,\ j \in [\mathrm{med}_+(n)]. \end{cases} \tag{26}$$

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

$$g_{n+1,i}(X_1^{N_{n+1}}, Y_1^{N_{n+1}}) = \begin{cases} \left(g_{n,j}(X_{N_n+1}^{2N_n}, Y_{N_n+1}^{2N_n}), g_{n,j+1}(X_1^{N_n}, Y_1^{N_n})\right), & i = 2j,\ j \in [\mathrm{med}_-(n)], \\ \left(g_{n,j+1}(X_1^{N_n}, Y_1^{N_n}), g_{n,j}(X_{N_n+1}^{2N_n}, Y_{N_n+1}^{2N_n})\right), & i = 2j,\ j \in [\mathrm{med}_+(n)], \\ \left(f_{n+1,i-1}(X_1^{N_{n+1}}), g_{n+1,i-1}(X_1^{N_{n+1}}, Y_1^{N_{n+1}})\right), & i = 2j+1. \end{cases} \tag{27}$$

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 $j \in [\mathrm{med}_-(n)]$ and $j \in [\mathrm{med}_+(n)]$ to ensure that, for even i, the first part of the observation is an observation from $[\mathrm{med}_-(n)]$ and the second part is an observation from $[\mathrm{med}_+(n)]$. This is an artifact of the medial 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 medial indices, because our analysis will focus on medial indices. The recursion (26), (27) is well-defined, as medial indices are only ever generated from medial indices (see Remark 3), so nowhere in the recursion will a non-medial index appear.

The *observation-truncated BST* is also a recursively-defined sequence of functions $\tilde{f}_{n,i}$ and $\tilde{g}_{n,i}$. 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 $i \in [\mathrm{med}(0)]$,

$$\tilde{f}_{0,i}(X_1^{N_0}) = X_i, \tag{28a}$$
$$\tilde{g}_{0,i}(X_1^{N_0}, Y_1^{N_0}) = (X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}). \tag{28b}$$

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

$$\tilde{g}_{n,i}(X_1^{N_n}, Y_1^{N_n}) = \gamma_{n,i}(g_{n,i}(X_1^{N_n}, Y_1^{N_n})) = \gamma_{n,i}(G_i).$$

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

The domains for $f_{n,i}, \tilde{f}_{n,i}, g_{n,i}, \tilde{g}_{n,i}$ 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 $\mathbf{i} = \begin{bmatrix} i_1 & i_2 & \cdots & i_k \end{bmatrix}$, natural numbers L, M, and a sequence of random variables X_j, we denote

$$X_{\mathbf{i}} = (X_{i_1}, X_{i_2}, \ldots, X_{i_k}), \tag{29a}$$
$$X_{\mathbf{i}-L} = (X_{i_1-L}, X_{i_2-L}, \ldots, X_{i_k-L}), \tag{29b}$$
$$X_{\mathbf{i}-L}^{\mathbf{i}+M} = (X_{i_1-L}^{i_1+M}, S_{i_2-L}^{i_2+M}, \ldots, X_{i_k-L}^{i_2+M}). \tag{29c}$$

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

$$f_{n,i}(X_1^{N_n}) = f_{n,i}(X_{\mathbf{b}}).$$

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

$$g_{n,i}(X_1^{N_n}, Y_1^{N_n}) = g_{n,i}(X_{\mathbf{a}}^{\mathbf{b}}, Y_{\mathbf{a}}^{\mathbf{z}}),$$
$$\tilde{g}_{n,i}(X_1^{N_n}, Y_1^{N_n}) = \tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}),$$

where we denoted

$$\mathbf{a} = \begin{bmatrix} 1 & N_0+1 & 2N_0+1 & \cdots & (2^n-1)N_0+1 \end{bmatrix},$$
$$\mathbf{z} = \begin{bmatrix} N_0 & 2N_0 & 3N_0 & \cdots & 2^n N_0 \end{bmatrix}.$$

Note that $Y_{\mathbf{a}}^{\mathbf{z}} = Y_1^{N_n}$.

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

$$\mathbf{b} = \begin{bmatrix} 6 & 17 & 29 & 40 & 53 & 64 & 77 & 88 \end{bmatrix}.$$

We have (see Figure 7):

$$F_{34} = f_{3,34}(X_{\mathbf{b}}) = X_6 + X_{17} + X_{40} + X_{77} + X_{88},$$
$$F_{35} = f_{3,35}(X_{\mathbf{b}}) = X_6 + X_{17} + X_{40}.$$

Recall that $\mathbf{b}$ is the base-vector of level-n index i. From the recursive definition (27), we observe that we can compute $X_{\mathbf{b}-L_0}^{\mathbf{b}-1}$ from $\tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0})$. This is easily shown by induction. It is trivially true for $n = 0$. Assume that this holds for $n \geq 0$ for any medial index; we will show it holds for $n+1$ as well. Indeed, write $\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 \end{bmatrix}$, where $\mathbf{b}_1$ and $\mathbf{b}_2$ are of length 2^{n-1}. By the recursive definition of $\mathbf{b}$, (22), the recursion (27) becomes

$$\tilde{g}_{n+1,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}) = \begin{cases} \left(\tilde{g}_{n,j}(X_{\mathbf{b}_2-L_0}^{\mathbf{b}_2}, Y_{\mathbf{b}_2-L_0}^{\mathbf{b}_2+L_0}), \tilde{g}_{n,j+1}(X_{\mathbf{b}_1-L_0}^{\mathbf{b}_1}, Y_{\mathbf{b}_1-L_0}^{\mathbf{b}_1+L_0})\right), & i=2j,\ j\in[\mathrm{med}_-(n)], \\ \left(\tilde{g}_{n,j+1}(X_{\mathbf{b}_1-L_0}^{\mathbf{b}_1}, Y_{\mathbf{b}_1-L_0}^{\mathbf{b}_1+L_0}), \tilde{g}_{n,j}(X_{\mathbf{b}_2-L_0}^{\mathbf{b}_2}, Y_{\mathbf{b}_2-L_0}^{\mathbf{b}_2+L_0})\right), & i=2j,\ j\in[\mathrm{med}_+(n)], \\ \left(\tilde{f}_{n+1,i-1}(X_{\mathbf{b}-L_0}^{\mathbf{b}}), \tilde{g}_{n+1,i-1}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0})\right), & i = 2j+1. \end{cases}$$

By the induction hypothesis, we can compute $X_{\mathbf{b}_1-L_0}^{\mathbf{b}_1-1}$ from $\tilde{g}_{n,j+1}(X_{\mathbf{b}_1-L_0}^{\mathbf{b}_1})$, and $X_{\mathbf{b}_2-L_0}^{\mathbf{b}_2-1}$ from $\tilde{g}_{n,j}(X_{\mathbf{b}_2-L_0}^{\mathbf{b}_2}, Y_{\mathbf{b}_2-L_0}^{\mathbf{b}_2+L_0})$. In other words, we can compute $X_{\mathbf{b}-L_0}^{\mathbf{b}-1}$ from $\tilde{g}_{n+1,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0})$. Of course, one can also compute $Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}$ from $\tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0})$. 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,

$$\tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}) \equiv \left(\tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}), X_{\mathbf{b}-L_0}^{\mathbf{b}-1}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}\right). \tag{30}$$

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

$$G_i = g_{n,i}(X_{\mathbf{a}}^{\mathbf{b}}, Y_{\mathbf{a}}^{\mathbf{z}}) \equiv (\acute{G}_i, \underset{\prime}{G}_i), \tag{31}$$

where

$$\begin{aligned} \acute{G}_i &= \tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}), \\ \underset{\prime}{G}_i &= (X_{\mathbf{a}}^{\mathbf{b}-L_0-1}, Y_{\mathbf{a}}^{\mathbf{b}-L_0-1}, Y_{\mathbf{b}+L_0+1}^{\mathbf{z}}). \end{aligned}$$

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 $\acute{G}_i, \underset{\prime}{G}_i$ rather than $\tilde{G}_i, \underset{\sim}{G}_i$. 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 $\mathbf{b}$ we call $\tilde{f}_{n,i}(X_{\mathbf{b}})$ an OT-transformed index. The conditional entropy of OT-transformed level-n index i is $H(\tilde{f}_{n,i}(X_{\mathbf{b}})|\tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}))$. 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 $H(\tilde{f}_{n,i}(X_{\mathbf{b}})|\tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0})) < \eta$ for all $i \in \mathcal{L}$, or $H(\tilde{f}_{n,i}(X_{\mathbf{b}})|\tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0})) > 1 - \eta$ 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 $X_\star \rightarrowtail Y_\star$ 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 $\mathcal{Y}, \mathcal{S}$ 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, for any any j,

$$P_{S_j, X_j, Y_j | S_{-\infty}^{j-1}, X_{-\infty}^{j-1}, Y_{-\infty}^{j-1}} = P_{S_j, X_j, Y_j | S_{j-1}} = P_{S_j | S_{j-1}} \cdot P_{X_j, Y_j | S_j}, \tag{32}$$

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

An s/o-process $X_\star \rightarrowtail Y_\star$ 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_k, X_k, Y_k are independent of S_{l-1}, X_l, Y_l, for any $l < j \leq k$. Furthermore, X_j, Y_j are a function (possibly probablistic) 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),[6] 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 $X_\star \rightarrowtail Y_\star$ is a FAIM-derived s/o-process, there exist sequences ψ_k, ϕ_k, $k \geq 0$, such that for any $L \leq M \in \mathbb{Z}$,*

$$P_{X_{-\infty}^{L}, Y_{-\infty}^{L}, X_{M+1}^{\infty}, Y_{M+1}^{\infty}} \leq \psi_{M-L} \cdot P_{X_{-\infty}^{L}, Y_{-\infty}^{L}} \cdot P_{X_{M+1}^{\infty}, Y_{M+1}^{\infty}}, \tag{33a}$$
$$P_{X_{-\infty}^{L}, Y_{-\infty}^{L}, X_{M+1}^{\infty}, Y_{M+1}^{\infty}} \geq \phi_{M-L} \cdot P_{X_{-\infty}^{L}, Y_{-\infty}^{L}} \cdot P_{X_{M+1}^{\infty}, Y_{M+1}^{\infty}}. \tag{33b}$$

The sequence ψ_k is nonincreasing and the sequence ϕ_k is nondecreasing. Both ψ_k and ϕ_k tend to 1 exponentially fast as $k \to \infty$.

The sequences ψ_k and ϕ_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

$$\begin{aligned} \psi_k &= \max_{s,\sigma} \frac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)\,\mathbb{P}(S_k = \sigma)}, \\ \phi_k &= \min_{s,\sigma} \frac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)\,\mathbb{P}(S_k = \sigma)} \end{aligned}$$

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: $\mathbb{P}(S_k = s) > 0$ for any $s \in \mathcal{S}$ and $k \in \mathbb{Z}$, [32, Theorem 4.2].

It is immediately evident that for any $k \geq 1$, $1 \leq \psi_k < \infty$ 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, ϕ_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"

[6]Indeed, the redefined Markov chain remains finite-state, aperiodic, and irreducible. The redefined state $\bar{S}_j$ takes values in alphabet $\bar{\mathcal{S}} = \{(s_j, s_{j-1}) \mid s_j, s_{j-1} \in \mathcal{S},\ P_{S_j|S_{j-1}}(s_j|s_{j-1}) > 0\}$. It assumes the value $\bar{S}_j = (s_j, s_{j-1})$ whenever $S_j = s_j, S_{j-1} = s_{j-1}$. Since $|\mathcal{S}| < \infty$, so is $|\bar{\mathcal{S}}| < \infty$. The original Markov chain is aperiodic and irreducible if and only if there exists $k > 0$ such that $P_{S_k|S_0}(s_k|s_0) > 0$ for any $s_0, s_k \in \mathcal{S}$. For this k and any $\bar{s}_0 = (s_0, s_{-1}) \in \bar{\mathcal{S}}$ and $\bar{s}_{k+1} = (s_{k+1}, s_k) \in \bar{\mathcal{S}}$, we have $P_{\bar{S}_{k+1}|\bar{S}_0}(\bar{s}_{k+1}|\bar{s}_0) > 0$. Thus, the redefined Markov process remains finite-state, aperiodic, and irreducible.

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 λ such that if $k \geq \lambda$ then

$$I(S_1; S_k | X_1^k, Y_1^k) \leq \epsilon, \tag{34a}$$
$$I(S_1; S_k | Y_1^k) \leq \epsilon. \tag{34b}$$

We call ϵ the *forgetfulness* of the s/o-process, and λ the *recollection* of the process. The recollection for a given ϵ is called 'ϵ-recollection.' The forgetfulness for a given λ is called 'λ-forgetfulness.'

We say that FAIM-derived s/o-process $X_\star \rightarrowtail Y_\star$ 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 ϵ, 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 $m \geq k$,

$$I(S_\ell; S_m | X_1^k, Y_1^k) \leq \epsilon, \tag{35a}$$
$$I(S_\ell; S_m | Y_1^k) \leq \epsilon. \tag{35b}$$

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

$$\mathsf{M} = \begin{bmatrix} 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix},$$

and the observation Y_j is given by

$$Y_j = \begin{cases} a, & \text{if } S_j \in \{1, 2\}, \\ b, & \text{if } S_j \in \{3, 4\}. \end{cases} \tag{36}$$

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, $\mathsf{M}^3 > 0$.

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.

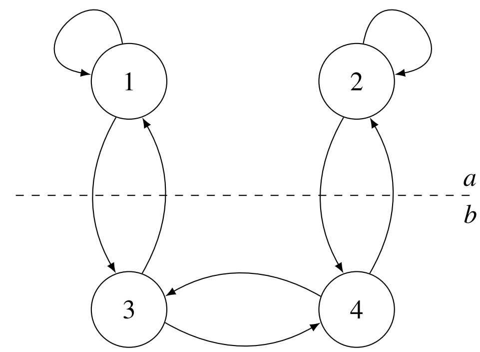

Fig. 8. The Markov chain S_j has four states. The possible transitions are depicted using arrows; the probability of choosing any transition is 1/2. The observation Y_j is 'a' if $S_j \in \{1, 2\}$ or 'b' if $S_j \in \{3, 4\}$.

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, $I(S_1; S_k | Y_1^k)$ 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, $I(S_1; S_k | X_1^k, Y_1^k)$ will vanish with k but $I(S_1; S_k | Y_1^k)$ will not. For the second process, $I(S_1; S_k | X_1^k, Y_1^k)$ will not vanish with k but $I(S_1; S_k | Y_1^k)$ will.

- Let $X_j = S_j$ and Y_j as in (36). Then, $I(S_1; S_k | X_1^k, Y_1^k) = I(S_1; S_k | S_1^k) = 0$ trivially. On the other hand, as shown in Example 3, $I(S_1; S_k | Y_1^k)$ 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, $I(S_1; S_k | X_1^k, Y_1^k)$ cannot vanish with k, as shown in Example 3. On the other hand, $I(S_1; S_k | Y_1^k) = I(S_1; S_k) \to 0$, 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 to it a level-0 BST, initialized with L_0 that is greater than its ϵ-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 λ such that for any integers m, ℓ, k such that $\min\{m, \ell\} \geq k \geq \lambda$ we have*

$$I(S_0; S_{-k}, S_k | X_{-\ell}^{-1}, Y_{-\ell}^m) \leq 2\epsilon. \tag{37}$$

This is a consequence of (34). To prove it, we take λ as the ϵ-recollection 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 λ 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

$$I(S_{\mathbf{i}}; S_{\mathbf{i}-L_0}, S_{\mathbf{i}+L_0} | X_{\mathbf{i}-L_0}^{\mathbf{i}-1}, Y_{\mathbf{i}-L_0}^{\mathbf{i}+L_0}) \leq k \cdot 2\epsilon,$$

where

$$\mathbf{i} = \begin{bmatrix} i_1 & i_2 & \cdots & i_k \end{bmatrix}.$$

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,

$$\tilde{\mathcal{H}} \triangleq H(X_i | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}). \tag{38}$$

By stationarity, $\tilde{\mathcal{H}}$ is indeed independent of i.

Lemma 9. *Let $X_\star \rightarrowtail Y_\star$ be a forgetful FAIM-derived s/o-process with ϵ-recollection λ. Let $L_0 \geq \lambda$ and $M_0 \geq 1$, and denote $N_0 = 2L_0 + M_0$. Then, for any $L_0 + 1 \leq i \leq L_0 + M_0$ we have*

$$0 \leq \tilde{\mathcal{H}} - H(X_i | X_1^{i-1}, Y_1^{N_0}) \leq 2\epsilon. \tag{39}$$

Proof: Observe that

$$\begin{aligned}
&\tilde{\mathcal{H}} - H(X_i | X_1^{i-1}, Y_1^{N_0}) \\
&\quad = H(X_i | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}) - H(X_i | X_1^{i-1}, Y_1^{N_0}) \\
&\quad = I\left(X_i; (X_1^{i-L_0-1}, Y_1^{i-L_0-1}, Y_{i+L_0+1}^{N_0}) | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}\right).
\end{aligned}$$

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

Let (S_j, X_j, Y_j) be the FAIM process from which $X_\star \rightarrowtail Y_\star$ is derived. By stationarity and Lemma 7, for any m, ℓ, k such that $\min\{m, \ell\} \geq k \geq \lambda$,

$$I(S_i; S_{i-k}, S_{i+k} | X_{i-\ell}^{i-1}, Y_{i-\ell}^{i+m}) \leq 2\epsilon. \tag{40}$$

Setting $\ell = m = k = L_0$ in (40) yields

$$I(S_i; S_{i-L_0}, S_{i+L_0} | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}) \leq 2\epsilon.$$

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

$$\begin{aligned}
2\epsilon &\geq I\left(S_i; S_{i-L_0}, S_{i+L_0} | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}\right) \\
&\overset{(a)}{\geq} I\left(X_i; S_{i-L_0}, S_{i+L_0} | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}\right) \\
&\overset{(b)}{\geq} I\left(X_i; X_1^{i-L_0-1}, Y_1^{i-L_0-1}, Y_{i+L_0+1}^{N_0} | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}\right).
\end{aligned}$$

We now detail the Markov chains used for the inequalities, both using (32). Inequality (a) is due to

$$(S_{i-L_0}, S_{i+L_0}) \multimap (S_i, X_{i-L_0}^{i-1}, Y_{i-L_0}^{i-1}) \multimap X_i,$$

and inequality (b) is due to

$$X_i \multimap (S_{i-L_0}, S_{i+L_0}, X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}) \multimap (X_1^{i-L_0-1}, Y_1^{i-L_0-1}, Y_{i+L_0+1}^{N_0}).$$

This completes the proof. ∎

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

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

$$\left|\mathcal{H}(X_\star | Y_\star) - \tilde{\mathcal{H}}\right| \leq 2\epsilon, \tag{41}$$

Proof: For any $\xi > 0$, let $N = N(\xi) > 2L_0$ be large enough so that $|\mathcal{H}(X_\star | Y_\star) - H(X_1^N | Y_1^N)/N| \leq \xi/2$ and $2L_0/N \leq \xi/2$. Then,

$$\begin{aligned}
&|\mathcal{H}(X_\star | Y_\star) - \tilde{\mathcal{H}}| \\
&\overset{(a)}{\leq} \left|\mathcal{H}(X_\star | Y_\star) - \frac{1}{N} H(X_1^N | Y_1^N)\right| + \left|\frac{1}{N} H(X_1^N | Y_1^N) - \tilde{\mathcal{H}}\right| \\
&\overset{(b)}{\leq} \frac{\xi}{2} + \frac{1}{N} \sum_{i=1}^{N} \left|H(X_i | X_1^{i-1}, Y_1^N) - \tilde{\mathcal{H}}\right| \\
&\overset{(c)}{\leq} \frac{\xi}{2} + \frac{2L_0}{N} + \frac{1}{N} \sum_{i=L_0+1}^{N-L_0} \left|H(X_i | X_1^{i-1}, Y_1^N) - \tilde{\mathcal{H}}\right| \\
&\overset{(d)}{\leq} \xi + \frac{N - 2L_0}{N} 2\epsilon \\
&\leq 2\epsilon + \xi,
\end{aligned}$$

where (a) and (b) are by the triangle inequality; (c) is because $|H(X_i | X_1^{i-1}, Y_1^N) - \tilde{\mathcal{H}}| \leq \max\{\tilde{\mathcal{H}}, H(X_i | X_1^{i-1}, Y_1^N)\} \leq 1$, 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 $N_n = 2^n N_0$, where $N_0 = 2L_0 + M_0$. Denote by $P_{X_1^{N_n}, Y_1^{N_n}}$ the joint distribution of $(X_1^{N_n}, Y_1^{N_n})$. By marginalizing $P_{X_1^{N_n}, Y_1^{N_n}}$, we obtain the distribution of a single b-block, $P_{X_{(\ell-1)N_0+1}^{\ell N_0}, Y_{(\ell-1)N_0+1}^{\ell N_0}}$, which, by stationarity, is independent of ℓ.

Definition 11 (Block-Independent Process)**.** The *block-independent process* (BI-process) $\tilde{X}_\star \rightarrowtail \tilde{Y}_\star$ with parameter N_0, is distributed according to

$$(\tilde{X}_1^{N_n}, \tilde{Y}_1^{N_n}) \sim \prod_{\ell=1}^{2^n} P_{X_{(\ell-1)N_0+1}^{\ell N_0}, Y_{(\ell-1)N_0+1}^{\ell N_0}}.$$

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

If $\mathbf{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_{2^n} \end{bmatrix}$ is the base-vector of a level-n medial index, we have

$$\tilde{X}_{\mathbf{b}-L_0}^{\mathbf{b}}, \tilde{Y}_{\mathbf{b}-L_0}^{\mathbf{b}+L_0} \sim \prod_{\ell=1}^{2^n} P_{X_{b_\ell - L_0}^{b_\ell}, Y_{b_\ell - L_0}^{b_\ell + L_0}}, \tag{42}$$

where $P_{X_{b_\ell - L_0}^{b_\ell}, Y_{b_\ell - L_0}^{b_\ell + L_0}}$ is obtained from $P_{X_{(\ell-1)N_0+1}^{\ell N_0}, Y_{(\ell-1)N_0+1}^{\ell N_0}}$ by marginalization. Note that since each b_ℓ is medial, $(X_{b_\ell - L_0}^{b_\ell}, Y_{b_\ell - L_0}^{b_\ell + L_0})$ is wholly contained in a b-block with b-block number ℓ.

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

$$\mathbf{a} = \begin{bmatrix} 1 & N_0+1 & 2N_0+1 & \cdots & (2^n-1)N_0+1 \end{bmatrix},$$
$$\mathbf{z} = \begin{bmatrix} N_0 & 2N_0 & 3N_0 & \cdots & 2^n N_0 \end{bmatrix}.$$

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

$$\tilde{F}_i = \tilde{f}_{n,i}(\tilde{X}_{\mathbf{b}}), \quad \tilde{G}_i = \tilde{g}_{n,i}(\tilde{X}_{\mathbf{b}-L_0}^{\mathbf{b}}, \tilde{Y}_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}). \tag{43a}$$
$$\tilde{F}_j = \tilde{f}_{n,j}(\tilde{X}_{\mathbf{d}}), \quad \tilde{G}_j = \tilde{g}_{n,j}(\tilde{X}_{\mathbf{d}-L_0}^{\mathbf{d}}, \tilde{Y}_{\mathbf{d}-L_0}^{\mathbf{d}+L_0}). \tag{43b}$$

The joint distribution of $(\tilde{X}_{\mathbf{b}-L_0}^{\mathbf{b}}, \tilde{Y}_{\mathbf{b}-L_0}^{\mathbf{b}+L_0})$ is given by (42) with $\mathbf{b}$ as the base-vector of i. The joint distribution of $(\tilde{X}_{\mathbf{d}-L_0}^{\mathbf{d}}, \tilde{Y}_{\mathbf{d}-L_0}^{\mathbf{d}+L_0})$ is given by (42) with $\mathbf{b}$ set to $\mathbf{d}$, the base-vector of j.

Recall from (38) that we denoted $\tilde{\mathcal{H}} = H(X_i | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0})$, which, by stationarity, is independent of i. We wish to show that there exists $\delta_n \geq 0$, independent of i, such that if $i \in [\mathrm{med}_-(n)]$ then $H(\tilde{F}_i|\tilde{G}_i) = \tilde{\mathcal{H}} + \delta_n$ and if $i \in [\mathrm{med}_+(n)]$ then $H(\tilde{F}_i|\tilde{G}_i) = \tilde{\mathcal{H}} - \delta_n$. This will follow as a corollary to the following lemma.

Lemma 11. *Suppose that either $i, j \in [\mathrm{med}_-(n)]$ or $i, j \in [\mathrm{med}_+(n)]$. Then, the joint distribution of $(\tilde{F}_i, \tilde{G}_i)$ is the same as the joint distribution of $(\tilde{F}_j, \tilde{G}_j)$.*

Proof: We use induction. For $n = 0$, the claim is true by stationarity and the initialization of the OT-BST, (28). Indeed, in this case, $\tilde{F}_i = \tilde{X}_i$, $\tilde{F}_j = \tilde{X}_j$, $\tilde{G}_i = (\tilde{X}_{i-L_0}^{i-1}, \tilde{Y}_{i-L_0}^{i+L_0})$, and $\tilde{G}_j = (\tilde{X}_{j-L_0}^{j-1}, \tilde{Y}_{j-L_0}^{j+L_0})$. Stationarity implies that the joint distribution of $(\tilde{F}_i, \tilde{G}_i)$ is the same as the joint distribution of $(\tilde{F}_j, \tilde{G}_j)$.

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

Denote $i' = \lfloor i/2 \rfloor$ and $j' = \lfloor j/2 \rfloor$. We write $\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 \end{bmatrix}$ and $\mathbf{d} = \begin{bmatrix} \mathbf{d}_1 & \mathbf{d}_2 \end{bmatrix}$, where $\mathbf{b}_1, \mathbf{b}_2, \mathbf{d}_1, \mathbf{d}_2$ are vectors of length 2^{n-1}. Then, $\mathbf{b}_1$ is the base-vector of $i' + 1$, and $\mathbf{b}_2$ is the base-vector of i', see (22). Similarly, $\mathbf{d}_1$ is the base-vector of $j' + 1$, and $\mathbf{d}_2$ is the base-vector of j'. Denote

$$\tilde{U}_{i'+1} = \tilde{f}_{n-1,i'+1}(\tilde{X}_{\mathbf{b}_1}), \ \tilde{Q}_{i'+1} = \tilde{g}_{n-1,i'+1}(\tilde{X}_{\mathbf{b}_1-L_0}^{\mathbf{b}_1}, \tilde{Y}_{\mathbf{b}_1-L_0}^{\mathbf{b}_1+L_0}), \tag{44a}$$
$$\tilde{V}_{i'} = \tilde{f}_{n-1,i'}(\tilde{X}_{\mathbf{b}_2}), \qquad \tilde{R}_{i'} = \tilde{g}_{n-1,i'}(\tilde{X}_{\mathbf{b}_2-L_0}^{\mathbf{b}_2}, \tilde{Y}_{\mathbf{b}_2-L_0}^{\mathbf{b}_2+L_0}). \tag{44b}$$

Of the two s/o-pairs $\tilde{U}_{i'+1} \rightarrowtail \tilde{Q}_{i'+1}$ and $\tilde{V}_{i'} \rightarrowtail \tilde{R}_{i'}$, one is in $[\mathrm{med}_-(n-1)]$ and the other in $[\mathrm{med}_+(n-1)]$. We denote by $\tilde{T}_i^-$ the pair that is in $[\mathrm{med}_-(n-1)]$ and by $\tilde{T}_i^+$ the pair that is in $[\mathrm{med}_+(n-1)]$. That is,

$$\tilde{T}_i^- = \begin{cases} (\tilde{V}_{i'}, \tilde{R}_{i'}), & i' \in [\mathrm{med}_-(n-1)], \\ (\tilde{U}_{i'+1}, \tilde{Q}_{i'+1}), & i' \in [\mathrm{med}_+(n-1)] \end{cases}$$

and

$$\tilde{T}_i^+ = \begin{cases} (\tilde{U}_{i'+1}, \tilde{Q}_{i'+1}), & i' \in [\mathrm{med}_-(n-1)], \\ (\tilde{V}_{i'}, \tilde{R}_{i'}), & i' \in [\mathrm{med}_+(n-1)]. \end{cases}$$

We similarly define $\tilde{U}_{j'+1}$, $\tilde{V}_{j'}$, $\tilde{Q}_{j'+1}$, $\tilde{R}_{j'}$, $\tilde{T}_j^-$, and $\tilde{T}_j^+$ (with $\mathbf{b}$ replaced with $\mathbf{d}$ and i' replaced with j').

For the BI-process, b-blocks are independent. In particular, by (42), $(\tilde{X}_{\mathbf{b}_1-L_0}^{\mathbf{b}_1}, \tilde{Y}_{\mathbf{b}_1-L_0}^{\mathbf{b}_1+L_0})$ is independent of $(\tilde{X}_{\mathbf{b}_2-L_0}^{\mathbf{b}_2}, \tilde{Y}_{\mathbf{b}_2-L_0}^{\mathbf{b}_2+L_0})$. Hence, $\tilde{T}_i^-$ and $\tilde{T}_i^+$ are independent. Similarly, $\tilde{T}_j^-$ and $\tilde{T}_j^+$ are independent. By the induction hypothesis, $\tilde{T}_i^-$ and $\tilde{T}_j^-$ have the same distribution; $\tilde{T}_i^+$ and $\tilde{T}_j^+$ are also equi-distributed. By block-independence, the joint distribution of $(\tilde{T}_i^-, \tilde{T}_i^+)$ is the same as the joint distribution of $(\tilde{T}_j^-, \tilde{T}_j^+)$.

Assume first that $i, j \in [\mathrm{med}_-(n)]$. We then have, by (26) and (27),

$$\tilde{F}_i = \tilde{U}_{i'+1} + \tilde{V}_{i'}, \quad \tilde{G}_i = \begin{cases} (\tilde{R}_{i'}, \tilde{Q}_{i'+1}), & i' \in [\mathrm{med}_-(n-1)], \\ (\tilde{Q}_{i'+1}, \tilde{R}_{i'}), & i' \in [\mathrm{med}_+(n-1)], \end{cases} \tag{45}$$

and

$$\tilde{F}_j = \tilde{U}_{j'+1} + \tilde{V}_{j'}, \quad \tilde{G}_j = \begin{cases} (\tilde{R}_{j'}, \tilde{Q}_{j'+1}), & j' \in [\mathrm{med}_-(n-1)], \\ (\tilde{Q}_{j'+1}, \tilde{R}_{j'}), & j' \in [\mathrm{med}_+(n-1)]. \end{cases} \tag{46}$$

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

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

$$\tilde{F}_i = \begin{cases} \tilde{V}_{i'}, & i' \in [\mathrm{med}_-(n-1)], \\ \tilde{U}_{i'+1}, & i' \in [\mathrm{med}_+(n-1)]. \end{cases}$$

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

By (26) and (27),

$$(\tilde{F}_i, \tilde{G}_i) = (\tilde{F}_i, \tilde{F}_{i-1}, \tilde{G}_{i-1}) \equiv (\tilde{F}_i, \tilde{F}_i + \tilde{F}_{i-1}, \tilde{G}_{i-1}), \tag{47}$$

Similarly,

$$(\tilde{F}_j, \tilde{G}_j) = (\tilde{F}_j, \tilde{F}_{j-1}, \tilde{G}_{j-1}) \equiv (\tilde{F}_j, \tilde{F}_j + \tilde{F}_{j-1}, \tilde{G}_{j-1}). \tag{48}$$

The mappings on the right-hand sides of (47) and (48) are the same. Moreover, by (27), the mapping between $(\tilde{F}_i, \tilde{F}_i +$

$\tilde{F}_{i-1}, \tilde{G}_{i-1})$ and $(\tilde{T}_i^-, \tilde{T}_i^+)$ is the same as the mapping between $(\tilde{F}_j, \tilde{F}_j + \tilde{F}_{j-1}, \tilde{G}_{j-1})$ and $(\tilde{T}_j^-, \tilde{T}_j^+)$. Since $(\tilde{T}_i^-, \tilde{T}_i^+)$ and $(\tilde{T}_j^-, \tilde{T}_j^+)$ are equi-distributed, so are $(\tilde{F}_i, \tilde{G}_i)$ and $(\tilde{F}_j, \tilde{G}_j)$. ∎

Corollary 12. *There exists a nondecreasing sequence $\delta_n \geq 0$, independent of i, such that if $i \in [\text{med}_-(n)]$ then $H(\tilde{F}_i|\tilde{G}_i) = \tilde{\mathcal{H}} + \delta_n$ and $H(\tilde{F}_{i+1}|\tilde{G}_{i+1}) = \tilde{\mathcal{H}} - \delta_n$.*

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

$$H(\tilde{F}_i|\tilde{G}_i) = \begin{cases} \tilde{\mathcal{H}} + \delta_n, & i \in [\text{med}_-(n)], \\ \tilde{\mathcal{H}} - \delta_n, & i \in [\text{med}_+(n)]. \end{cases} \tag{49}$$

Proof: We show this using induction. The claim is true for $n = 0$ with $\delta_0 = 0$. For $n > 0$, 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 $\tilde{F}_i, \tilde{G}_i$, as well as $\tilde{F}_{i+1}, \tilde{G}_{i+1}$, be defined as in (43a) and let $\tilde{U}_{i'+1}, \tilde{V}_{i'}, \tilde{Q}_{i'+1}, \tilde{R}_{i'}$ be defined as in (44). We have, by (26) and (27),

$$\begin{aligned} H(\tilde{F}_i|\tilde{G}_i) + H(\tilde{F}_{i+1}|\tilde{G}_{i+1}) &= H(\tilde{F}_i, \tilde{F}_{i+1}|\tilde{Q}_{i'+1}, \tilde{R}_{i'}) \\ &= H(\tilde{U}_{i'+1}, \tilde{V}_{i'}|\tilde{Q}_{i'+1}, \tilde{R}_{i'}) \\ &= H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1}) + H(\tilde{V}_{i'}|\tilde{R}_{i'}), \end{aligned} \tag{50}$$

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

$$H(\tilde{U}_{i'}|\tilde{Q}_{i'}) = H(\tilde{V}_{i'}|\tilde{R}_{i'}) = \begin{cases} \tilde{\mathcal{H}} + \delta_{n-1}, & i' \in [\text{med}_-(n-1)], \\ \tilde{\mathcal{H}} - \delta_{n-1}, & i' \in [\text{med}_+(n-1)]. \end{cases}$$

Thus,

$$H(\tilde{F}_i|\tilde{G}_i) + H(\tilde{F}_{i+1}|\tilde{G}_{i+1}) = 2\tilde{\mathcal{H}}. \tag{51}$$

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

$$\begin{aligned} H(\tilde{F}_{i+1}|\tilde{G}_{i+1}) &\leq \min\{H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1}), H(\tilde{V}_{i'+1}|\tilde{R}_{i'+1})\} \\ &= \tilde{\mathcal{H}} - \delta_{n-1}. \end{aligned} \tag{52}$$

From (51) and (52), we conclude that there must exist $\delta_n \geq \delta_{n-1} \geq 0$ such that $H(\tilde{F}_i|\tilde{G}_i) = \tilde{\mathcal{H}} + \delta_n$ and $H(\tilde{F}_{i+1}|\tilde{G}_{i+1}) = \tilde{\mathcal{H}} - \delta_n$. Finally, by Lemma 11, δ_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

$$\alpha * \beta = \alpha + \beta(1-2\alpha) = \beta + \alpha(1-2\beta), \tag{53}$$

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 $0 \leq \alpha_a, \beta_b \leq 1/2$, $a, b = 1, 2, \ldots, k$ 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 $\xi_1, \xi_2 > 0$,*

$$\sum_{a=1}^k p_a h_2(\alpha_a) \geq \xi_1, \quad \sum_{b=1}^k q_b h_2(\beta_b) \leq \xi_2, \tag{54}$$

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

$$\sum_{a=1}^k \sum_{b=1}^k p_a q_b \left(h_2(\alpha_a * \beta_b) - h_2(\beta_b)\right) \geq \Delta(\xi_1, \xi_2).$$

Recall that $i \in [\text{med}(n)]$, with base-vector $\mathbf{b} = \begin{bmatrix}\mathbf{b}_1 & \mathbf{b}_2\end{bmatrix}$, where $\mathbf{b}_1$ and $\mathbf{b}_2$ are of length 2^{n-1}. Assume further that $i \in [\text{med}_-(n)]$, so that i is even, and $i' = i/2$. We define $\tilde{F}_i, \tilde{G}_i$ as in (43a), and $\tilde{U}_{i'+1}, \tilde{V}_{i'}, \tilde{Q}_{i'+1}, \tilde{R}_{i'}$ as in (44).

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

$$H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1}), H(\tilde{V}_{i'}|\tilde{R}_{i'}) \in (\xi, 1-\xi) \tag{55}$$

then

$$H(\tilde{F}_i|\tilde{G}_i) - \max\{H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1}), H(\tilde{V}_{i'}|\tilde{R}_{i'})\} \geq \Delta(\xi, 1-\xi).$$

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 $H(\tilde{V}_{i'}|\tilde{R}_{i'}) \geq H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1})$. The proof of the other case is similar and omitted.

We will use the simplified notation

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

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

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

We also introduce the shorthand

$$\begin{aligned} \alpha_q &= \min_u \mathbb{P}\left(\tilde{U}_{i'+1} = u|\tilde{Q}_{i'+1} = q\right) = \min_u \tilde{p}(u|q), \\ \beta_r &= \min_v \mathbb{P}\left(\tilde{V}_{i'} = v|\tilde{R}_{i'} = r\right) = \min_v \tilde{p}(v|r). \end{aligned}$$

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

$$\begin{aligned} H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1}) &= \sum_q \tilde{p}(q) h_2(\alpha_q) \geq \xi, \\ H(\tilde{V}_{i'}|\tilde{R}_{i'}) &= \sum_r \tilde{p}(r) h_2(\alpha_r) \leq 1-\xi. \end{aligned}$$

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

$$\begin{aligned} H(\tilde{F}_i|\tilde{G}_i) - H(\tilde{V}_{i'}|\tilde{R}_{i'}) &= H(\tilde{U}_{i'+1} + \tilde{V}_{i'}|\tilde{Q}_{i'+1}, \tilde{R}_{i'}) - H(\tilde{V}_{i'}|\tilde{R}_{i'}) \\ &= \sum_{q,r} \tilde{p}(q)\tilde{p}(r)\left(h_2(\alpha_q * \beta_r) - h_2(\beta_r)\right) \\ &\geq \Delta(\xi, 1-\xi), \end{aligned}$$

where the inequality is by Lemma 13. ∎

We are now ready to show that the OT-BST is monopolarizing for the BI-process. To this end, recall that $\tilde{\mathcal{H}}$ was defined in (38).

Proposition 15. *For every $\xi > 0$, there exists a threshold value $n_{\text{th}} \geq 0$ such that if $n \geq n_{\text{th}}$ then a level-n OT-BST with any parameters L_0, M_0 is $(\xi, [\text{med}_+(n)], [\text{med}_-(n)])$-monopolarizing for BI-process $\tilde{X}_\star \rightarrowtail \tilde{Y}_\star$ with parameter $N_0 = 2L_0 + M_0$.*

Specifically, let $\tilde{F}_1^{N_n} \rightarrowtail \tilde{G}_1^{N_n}$ be an OT-transformed s/o-block of a level-n OT-BST initialized with L_0 and M_0 as above, where $n \geq n_{\text{th}}$. Then:

- *if $\tilde{\mathcal{H}} \leq 1/2$ then $H(\tilde{F}_i|\tilde{G}_i) < \xi$, $\quad \forall i \in [\text{med}_+(n)]$;*
- *if $\tilde{\mathcal{H}} \geq 1/2$ then $H(\tilde{F}_i|\tilde{G}_i) > 1 - \xi$, $\quad \forall i \in [\text{med}_-(n)]$.*

Proof: Denote the indicator functions

$$\mathcal{M}_n^- = \begin{cases} 1, & H(\tilde{F}_i|\tilde{G}_i) > 1 - \xi, \quad \forall i \in [\text{med}_-(n)], \\ 0, & \text{otherwise}, \end{cases}$$

$$\mathcal{M}_n^+ = \begin{cases} 1, & H(\tilde{F}_i|\tilde{G}_i) < \xi, \quad \forall i \in [\text{med}_+(n)], \\ 0, & \text{otherwise}, \end{cases}$$

and

$$\mathcal{M}_n = \begin{cases} 1, & \mathcal{M}_n^- = 1 \text{ or } \mathcal{M}_n^+ = 1, \\ 0, & \text{otherwise}. \end{cases}$$

Observe that $\mathcal{M}_n = 1$ if and only if the OT-BST has $(\xi, [\text{med}_+(n)], [\text{med}_-(n)])$-monopolarized for the BI-process. Further define

$$n_{\text{th}} = \min\left\{n \in \mathbb{N} \mid \mathcal{M}_n = 1\right\}.$$

This is the first index n for which $\mathcal{M}_n = 1$. We will show that n_{th} is finite by upper-bounding it.

By Corollary 12, there exists a nondecreasing sequence $\delta_n \geq 0$ such that (49) holds. Since δ_n is a nondecreasing sequence, $\mathcal{M}_n = 1$ for every $n \geq n_{\text{th}}$. The entropy of a binary random variable is bounded between 0 and 1; thus for any n, $0 \leq \tilde{\mathcal{H}} - \delta_n \leq \tilde{\mathcal{H}} + \delta_n \leq 1$. Hence, $\delta_n \leq \min\{\tilde{\mathcal{H}}, 1 - \tilde{\mathcal{H}}\}$. We conclude that if $\tilde{\mathcal{H}} \leq 1/2$ and $n \geq n_{\text{th}}$ then $\mathcal{M}_n^+ = 1$, and if $\tilde{\mathcal{H}} \leq 1/2$ and $n \geq n_{\text{th}}$ then $\mathcal{M}_n^- = 1$. It now remains to upper-bound n_{th}.

If $\mathcal{M}_0 = 1$, then we may take $n_{\text{th}} = 0$ and we are done. Otherwise, we assume that $\mathcal{M}_0 = 0$.

If, for some $n \geq 0$, $\mathcal{M}_n = 0$, then by (49) and by definition of $\mathcal{M}_n^-$, $\mathcal{M}_n^+$, we obtain

$$\xi \leq \tilde{\mathcal{H}} - \delta_n \leq \tilde{\mathcal{H}} + \delta_n \leq 1 - \xi.$$

Rearranging, this yields

$$\mathcal{M}_n = 0 \Rightarrow \delta_n \leq \min\{\tilde{\mathcal{H}}, 1 - \tilde{\mathcal{H}}\} - \xi. \tag{56}$$

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

$$\tilde{\mathcal{H}} + \delta_n - (\tilde{\mathcal{H}} + \delta_{n-1}) \geq \Delta(\xi, 1 - \xi) \Rightarrow \delta_n \geq \delta_{n-1} + \Delta(\xi, 1 - \xi).$$

Continuing in this manner and recalling that $\delta_0 = 0$, we obtain

$$\mathcal{M}_{n-1} = 0 \Rightarrow \delta_n \geq n\Delta(\xi, 1 - \xi). \tag{57}$$

Now, let

$$n_1 = 1 + \left\lfloor \frac{\min\{\tilde{\mathcal{H}}, 1 - \tilde{\mathcal{H}}\} - \xi}{\Delta(\xi, 1 - \xi)} \right\rfloor, \tag{58}$$

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

$$n_1\Delta(\xi, 1 - \xi) \leq \delta_{n_1} \leq \min\{\tilde{\mathcal{H}}, 1 - \tilde{\mathcal{H}}\} - \xi.$$

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

$$n_1 \leq \frac{\min\{\tilde{\mathcal{H}}, 1 - \tilde{\mathcal{H}}\} - \xi}{\Delta(\xi, 1 - \xi)},$$

which contradicts (58) (see, e.g., [31, Equation 3.3]). We conclude that we must have $n_{\text{th}} \leq n_1$. 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 $0 \leq \zeta \leq 1$ and $n \geq n_{\text{th}}$ we have*

- *if $\tilde{\mathcal{H}} \leq \frac{1+\zeta}{2}$ then $H(\tilde{F}_i|\tilde{G}_i) < \xi + \zeta$, $\quad \forall i \in [\text{med}_+(n)]$,*
- *if $\tilde{\mathcal{H}} \geq \frac{1-\zeta}{2}$ then $H(\tilde{F}_i|\tilde{G}_i) > 1 - \xi - \zeta$, $\quad \forall i \in [\text{med}_-(n)]$.*

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

We only prove the corollary for the case where $\tilde{\mathcal{H}} \leq (1+\zeta)/2$. The case $\tilde{\mathcal{H}} \geq (1 - \zeta)/2$ is similar and omitted.

If $\tilde{\mathcal{H}} \leq 1/2$, we are done by Proposition 15. Otherwise, $\tilde{\mathcal{H}} \geq 1/2$, so by Proposition 15 and (49),

$$i \in [\text{med}_-(n)] \Rightarrow H(\tilde{F}_i|\tilde{G}_i) = \tilde{\mathcal{H}} + \delta_n > 1 - \xi.$$

Rearranging, we obtain $\delta_n > 1 - \tilde{\mathcal{H}} - \xi$. Now, by (49),

$$\begin{aligned} i \in [\text{med}_+(n)] \Rightarrow H(\tilde{F}_i|\tilde{G}_i) = \tilde{\mathcal{H}} - \delta_n &< \tilde{\mathcal{H}} - (1 - \tilde{\mathcal{H}} - \xi) \\ &= \xi + 2\tilde{\mathcal{H}} - 1 \\ &\leq \xi + (1 + \zeta) - 1 \\ &= \xi + \zeta, \end{aligned}$$

where the final inequality is due to our assumption that $\tilde{\mathcal{H}} \leq (1 + \zeta)/2$. ■

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

$$\begin{aligned} c(x, y) &= h_2(h_2^{-1}(y + x) * h_2^{-1}(y - x)) - y, \\ d(x, y) &= y - (y + x)(y - x), \end{aligned}$$

where $h_2^{-1} : [0, 1] \to [0, 1/2]$ is the inverse of h_2. Since h_2 is concave-∩ and increasing over $[0, 1/2]$, h_2^{-1} is convex-∪ and increasing over $[0, 1]$. We also define the sequence of functions

$$\begin{aligned} C_0(y) &= D_0(y) = 0, \\ C_n(y) &= c(C_{n-1}(y), y), \qquad n = 1, 2, \ldots, \\ D_n(y) &= d(D_{n-1}(y), y), \qquad n = 1, 2, \ldots. \end{aligned}$$

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

$$C_n(\tilde{\mathcal{H}}) \leq H(\tilde{F}_i|\tilde{G}_i) - \tilde{\mathcal{H}} \leq D_n(\tilde{\mathcal{H}}).$$

If $i \in [\mathrm{med}_+(n)]$ *then*

$$C_n(\tilde{\mathcal{H}}) \le \tilde{\mathcal{H}} - H(\tilde{F}_i|\tilde{G}_i) \le D_n(\tilde{\mathcal{H}}).$$

Proof: In light of Corollary 12, denote, for any $n \ge 0$ and arbitrary $i \in [\mathrm{med}_-(n)]$

$$\delta_n = H(\tilde{F}_i|\tilde{G}_i) - \tilde{\mathcal{H}}.$$

Observe that for arbitrary $i \in [\mathrm{med}_+(n)]$, by Corollary 12 we have $\delta_n = \tilde{\mathcal{H}} - H(\tilde{F}_i|\tilde{G}_i)$. Our goal is thus to show that for any $n \ge 0$,

$$C_n(\tilde{\mathcal{H}}) \le \delta_n \le D_n(\tilde{\mathcal{H}}). \tag{59}$$

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 $n \ge 0$, 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

$$\begin{aligned}\tilde{\mathcal{H}} + \delta_{n+1} &\ge h_2(h_2^{-1}(\tilde{\mathcal{H}} + \delta_n) * h_2^{-1}(\tilde{\mathcal{H}} - \delta_n)),\\ \tilde{\mathcal{H}} + \delta_{n+1} &\le (\tilde{\mathcal{H}} + \delta_n) + (\tilde{\mathcal{H}} - \delta_n) - (\tilde{\mathcal{H}} + \delta_n)(\tilde{\mathcal{H}} - \delta_n).\end{aligned}$$

Rearranging, we obtain

$$c(\delta_n, \tilde{\mathcal{H}}) \le \delta_{n+1} \le d(\delta_n, \tilde{\mathcal{H}}). \tag{60}$$

Now, $d(x, y) = x^2 - y^2 + y$ is increasing in x whenever $x \ge 0$. The function $c(x, y)$ is also increasing for $x \in [0, \min\{y, 1-y\}]$. To see this, it suffices to show that $c_y(x) = h_2^{-1}(y+x) * h_2^{-1}(y-x)$ is increasing, as h_2 is increasing. Denoting $r(x) = h_2^{-1}(x)$ we obtain that

$$\begin{aligned}\frac{\mathrm{d}c_y(x)}{\mathrm{d}x} &= r'(y+x)(1-2r(y-x)) - r'(y-x)(1-2r(y+x))\\ &\overset{(a)}{\ge} (r'(y+x) - r'(y-x))(1-2r(y+x))\\ &\overset{(b)}{\ge} 0,\end{aligned}$$

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 $r(\cdot) \le 1/2$ by definition. Thus, by (60) and the induction hypothesis (59),

$$\begin{aligned}\delta_{n+1} &\ge c(\delta_n, \tilde{\mathcal{H}}) \ge c(C_n(\tilde{\mathcal{H}}), \tilde{\mathcal{H}}) = C_{n+1}(\tilde{\mathcal{H}}),\\ \delta_{n+1} &\le d(\delta_n, \tilde{\mathcal{H}}) \le d(D_n(\tilde{\mathcal{H}}), \tilde{\mathcal{H}}) = D_{n+1}(\tilde{\mathcal{H}}),\end{aligned}$$

which completes the proof. ∎

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

Proposition 15 gives the upper bound

$$n_{\mathrm{th}} \le 1 + \left\lfloor \frac{\tilde{\mathcal{H}} - \xi}{\Delta(\xi, 1-\xi)} \right\rfloor = 40162.$$

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

$$\begin{aligned}2.22 \cdot 10^{-5} &\le H(\tilde{F}_i|\tilde{G}_i) \le 0.0041, &&i \in [\mathrm{med}_+(9)],\\ 8.89 \cdot 10^{-6} &\le H(\tilde{F}_i|\tilde{G}_i) \le 0.0031, &&i \in [\mathrm{med}_+(10)].\end{aligned}$$

Therefore, when $\tilde{\mathcal{H}} = 0.2$, $n_{\mathrm{th}} = 10$ suffices to ensure $(0.004, [\mathrm{med}_+(n)], [\mathrm{med}_-(n)])$-monopolarization for $n \ge n_{\mathrm{th}}$.

C. Monopolarization for FAIM-derived Processes

We now show that the BST is monopolarizing for suitably chosen η, $\mathcal{L}$, $\mathcal{H}$ when applied to forgetful FAIM-derived s/o-processes. Our main goal is to establish Theorem 18 below.

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

Specifically, let $F_1^{N_n} \rightarrowtail G_1^{N_n}$ be a transformed s/o-block of a level-n BST initialized with L_0 and M_0 as above. Then:

- *if $\mathcal{H}(X_\star|Y_\star) \le 1/2$ then $H(F_i|G_i) < \eta$, $\forall i \in [\mathrm{med}_+(n)]$;*
- *if $\mathcal{H}(X_\star|Y_\star) \ge 1/2$ then $H(F_i|G_i) > 1-\eta$, $\forall i \in [\mathrm{med}_-(n)]$.*

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 $X_\star \rightarrowtail Y_\star$ is a forgetful FAIM-derived s/o-process. By Lemma 6, it satisfies (33) with mixing sequences ψ_k, ϕ_k. We apply to s/o-block $X_1^{N_n} \rightarrowtail Y_1^{N_n}$ 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 $\tilde{X}_\star \rightarrowtail \tilde{Y}_\star$ 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 $i \in [\mathrm{med}(n)]$. The OT-BST is expressed using the sequence of functions $\tilde{f}_{n,i}$, $\tilde{g}_{n,i}$.

Let $i \in [\mathrm{med}(n)]$; its base-vector $\mathbf{b}$ is given by

$$\mathbf{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_{2^n} \end{bmatrix},$$

We also denote

$$\begin{aligned}\mathbf{a} &= \begin{bmatrix} 1 & N_0+1 & 2N_0+1 & \cdots & (2^n-1)N_0+1 \end{bmatrix},\\ \mathbf{z} &= \begin{bmatrix} N_0 & 2N_0 & 3N_0 & \cdots & 2^n N_0 \end{bmatrix}.\end{aligned}$$

We further define for index $i \in [\mathrm{med}(n)]$:

$$F_i = f_{n,i}(X_{\mathbf{b}}), \quad G_i = g_{n,i}(X_{\mathbf{a}}^{\mathbf{b}}, Y_{\mathbf{a}}^{\mathbf{z}}), \tag{61a}$$

$$\acute{F}_i = \tilde{f}_{n,i}(X_{\mathbf{b}}), \quad \acute{G}_i = \tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}), \tag{61b}$$

$$\tilde{F}_i = \tilde{f}_{n,i}(\tilde{X}_{\mathbf{b}}), \quad \tilde{G}_i = \tilde{g}_{n,i}(\tilde{X}_{\mathbf{b}-L_0}^{\mathbf{b}}, \tilde{Y}_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}). \tag{61c}$$

In words:

- $F_i \rightarrowtail G_i$ is a transformed s/o-pair obtained after applying a level-n BST to the FAIM-derived process;
- $\acute{F}_i \rightarrowtail \acute{G}_i$ is an OT-transformed s/o-pair obtained after applying a level-n OT-BST to the FAIM-derived process;
- $\tilde{F}_i \rightarrowtail \tilde{G}_i$ is an OT-transformed s/o-pair obtained after applying a level-n OT-BST to the BI-process.

Proposition 19. *Fix $n \ge 0$, $\varepsilon_1 > 0$, and $0 < \varepsilon_2 < \frac{1}{6}$. There exist L and M such that a level-n BST initialized with parameters $L_0 \ge L$, $M_0 \ge M$ satisfies:*

$$|H(F_i|G_i) - H(\tilde{F}_i|\tilde{G}_i)| < 2\varepsilon_1 + \sqrt{8\varepsilon_2}. \tag{62}$$

Proof: Denote

$$\acute{P} = P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}}\,,$$

$$\tilde{P} = \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}\,.$$

Then, $(X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0})$ is distributed according to $\acute{P}$ and $(\tilde{X}^{\mathbf{b}}_{\mathbf{b}-L_0}, \tilde{Y}^{\mathbf{b}+L_0}_{\mathbf{b}-L_0})$ is distributed according to $\tilde{P}$.

In Lemma 20 that follows we show that there exists L such that if $L_0 \geq L$ then

$$|H(F_i|G_i) - H(\acute{F}_i|\acute{G}_i)| \leq 2\varepsilon_1.$$

Next, in Lemma 21 that follows we show that there exists M such that if $M_0 \geq M$ then

$$(1-\varepsilon_2)\tilde{P} \leq \acute{P} \leq (1+\varepsilon_2)\tilde{P}.$$

This will enable us to use Lemma 22 below with $f = \tilde{f}_{n,i}$ and $g = \tilde{g}_{n,i}$ to obtain

$$|H(\acute{F}_i|\acute{G}_i) - H(\tilde{F}_i|\tilde{G}_i)| < \sqrt{8\varepsilon_2}.$$

Hence, we conclude that

$$\begin{aligned} &|H(F_i|G_i) - H(\tilde{F}_i|\tilde{G}_i)| \\ &\quad \leq |H(F_i|G_i) - F(\acute{F}_i|\acute{G}_i)| + |H(\acute{F}_i|\acute{G}_i) - H(\tilde{F}_i|\tilde{G}_i)| \\ &\quad < 2\varepsilon_1 + \sqrt{8\varepsilon_2}, \end{aligned}$$

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*

$$0 \leq H(\acute{F}_i|\acute{G}_i) - H(F_i|G_i) \leq 2\varepsilon_1. \tag{63}$$

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), $G_i \equiv (\acute{G}_i, \underset{,}{G}_i)$, where

$$\begin{aligned} \acute{G}_i &= \tilde{g}_{n,i}(X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}), \\ \underset{,}{G}_i &= (X^{\mathbf{b}-L_0-1}_{\mathbf{a}}, Y^{\mathbf{b}-L_0-1}_{\mathbf{a}}, Y^{\mathbf{z}}_{\mathbf{b}+L_0+1}). \end{aligned} \tag{64}$$

Since $f_{n,i} = \tilde{f}_{n,i}$, we have $F_i = \acute{F}_i$. Therefore,

$$H(F_i|G_i) = H(\acute{F}_i|\acute{G}_i, \underset{,}{G}_i) \leq H(\acute{F}_i|\acute{G}_i),$$

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 ϵ be the L-forgetfulness of the s/o-process; we soon specify how to set L. Now, utilize Corollary 8 with $\mathbf{i} = \mathbf{b}$, $\lambda = L$, and $L_0 \geq L$, to obtain

$$I(S_{\mathbf{b}}; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \leq 2^n \cdot 2\epsilon.$$

We take L large enough so that $\epsilon \leq \varepsilon_1 \cdot 2^{-n}$. Hence,

$$\begin{aligned} 2\varepsilon_1 &\geq I(S_{\mathbf{b}}; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \\ &\overset{(a)}{\geq} I(\acute{F}_i, \acute{G}_i; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \\ &\overset{(b)}{\geq} I(\acute{F}_i; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid \acute{G}_i, X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \\ &\overset{(c)}{=} I(\acute{F}_i; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid \acute{G}_i) \\ &= H(\acute{F}_i|\acute{G}_i) - H(\acute{F}_i|\acute{G}_i, S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) \\ &\overset{(d)}{=} H(\acute{F}_i|\acute{G}_i) - H(\acute{F}_i|\acute{G}_i, \underset{,}{G}_i, S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) \\ &\overset{(e)}{=} H(\acute{F}_i|\acute{G}_i) - H(F_i|G_i, S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) \\ &\overset{(f)}{\geq} H(\acute{F}_i|\acute{G}_i) - H(F_i|G_i), \end{aligned}$$

where:

- (a) is due to (2). By (32), $X_{\mathbf{b}}$ is a probabilistic function of $S_{\mathbf{b}}$; by (61b), $\acute{F}_i$ is a function of $X_{\mathbf{b}}$, and $\acute{G}_i$ is a function of $(X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0})$. Thus, we have the Markov chain
$$\begin{aligned} (S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) &-\!\!\circ\!\!- (S_{\mathbf{b}}, X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \\ &-\!\!\circ\!\!- (X_{\mathbf{b}}, X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) -\!\!\circ\!\!- (\acute{F}_i, \acute{G}_i). \end{aligned}$$
Specifically, we have the Markov chain
$$(S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) -\!\!\circ\!\!- (S_{\mathbf{b}}, X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) -\!\!\circ\!\!- (\acute{F}_i, \acute{G}_i),$$
for which we use (2).
- (b) is by the chain rule.
- (c) is since $\acute{G}_i \equiv (\acute{G}_i, X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0})$, which holds due to (30) and (61b).
- (d) is by the Markov property (32), (61b), and (64): $\acute{F}_i$ and $\acute{G}_i$ are probabilistic functions of states $S^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}$, whereas $\underset{,}{G}_i$ is a probabilistic function of states $S^{\mathbf{b}-L_0-1}_{\mathbf{a}}$ and $S^{\mathbf{z}}_{\mathbf{b}+L_0+1}$.
- (e) is because $\acute{F}_i = F_i$ and because $G_i \equiv (\acute{G}_i, \underset{,}{G}_i)$ by (31).
- (f) is because conditioning reduces entropy.

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*

$$P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} \leq (1+\varepsilon_2) \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}, \tag{65a}$$

$$P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} \geq (1-\varepsilon_2) \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}. \tag{65b}$$

Proof: Recall that the mixing sequences of the original s/o-process $X_\star \rightarrowtail Y_\star$ are ψ_k and ϕ_k, where $\psi_k \geq 1$ is nonincreasing and $\phi_k \leq 1$ is nondecreasing. By Lemma 6, both sequences approach 1 exponentially fast. Thus, we may choose M such that

$$\begin{aligned} (\psi_{M-2})^{(2^n)} &\leq 1 + \varepsilon_2, \\ (\phi_{M-2})^{(2^n)} &\geq 1 - \varepsilon_2. \end{aligned}$$

For any $M_0 \geq M$ we thus have

$$(\psi_{M_0-2})^{(2^n)} \leq (\psi_{M-2})^{(2^n)} \leq 1 + \varepsilon_2, \tag{66a}$$

$$(\phi_{M_0-2})^{(2^n)} \geq (\phi_{M-2})^{(2^n)} \geq 1 - \varepsilon_2. \tag{66b}$$

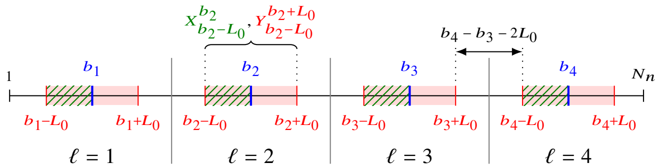

Fig. 9. Illustration of a level-2 BST. There are four b-blocks, with b-block numbers $\ell = 1, 2, 3, 4$. The base-index (in blue) in b-block ℓ is b_ℓ. The red boxes in the illustration correspond to $X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}$, where X is only available to the left of the blue lines (shown in green). Each red box represents a contiguous set of indices, and there are 2^n such sets; they are separated in time.

Denote by $\bar{\mathbf{b}} = \begin{bmatrix} \bar{b}_1 & \bar{b}_2 & \cdots & \bar{b}_{2^n} \end{bmatrix}$ the modulo-base-vector of i. By Corollary 5, for any $1 \le \ell < 2^n$ we have $1 \le |\bar{b}_{\ell+1} - \bar{b}_\ell| \le 2$. Hence, by (21), and recalling that $N_0 = 2L_0 + M_0$,

$$\begin{aligned}(b_{\ell+1} - L_0) - (b_\ell + L_0) &= \ell N_0 - (\ell-1)N_0 - 2L_0 + \bar{b}_{\ell+1} - \bar{b}_\ell \\ &= M_0 + (\bar{b}_{\ell+1} - \bar{b}_\ell) \\ &\ge M_0 - |\bar{b}_{\ell+1} - \bar{b}_\ell| \\ &\ge M_0 - 2. \end{aligned} \tag{67}$$

The vector $X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}$ contains symbols with indices in $\mathcal{B} = \cup_\ell \mathcal{B}_\ell$, where $\mathcal{B}_\ell = \{b_\ell - L_0, b_\ell - L_0 + 1, \ldots, b_\ell + L_0\}$, $1 \le \ell \le 2^n$. Each set $\mathcal{B}_\ell$ is a contiguous subsequence of $\mathcal{B}$. The greatest index in $\mathcal{B}_\ell$ is $b_\ell + L_0$ and the smallest index in $\mathcal{B}_{\ell+1}$ is $b_{\ell+1} - L_0$; see Figure 9 for an illustration. By (67), any two consecutive sets $\mathcal{B}_\ell$ and $\mathcal{B}_{\ell+1}$ are separated by at least $M_0 - 2$ indices.

Using Lemma 6 and a straightforward induction argument, we conclude that

$$\begin{aligned} P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} &\le (\psi_{M_0-2})^{(2^n)} \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}, \\ P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} &\ge (\phi_{M_0-2})^{(2^n)} \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}. \end{aligned}$$

Thus, by (66),

$$\begin{aligned} P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} &\le (1+\varepsilon_2) \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}, \\ P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} &\ge (1-\varepsilon_2) \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}, \end{aligned}$$

which is (65). ■

Remark 6. In the proof of Lemma 21 we saw that the parameter M is dependent on the convergence rate of the mixing sequences ψ_k, ϕ_k. Bounds on the convergence rates of these sequences that depend only on ψ_0 and the second-largest eigenvalue of a simple function of the transition matrix of the underlying Markov chain exist in the literature, see, e.g., [33, Theorem 2.1].

In Lemma 21 we saw that $\acute{P}$ and $\tilde{P}$ are close in the sense of (65). The following lemma, whose proof can be found in Appendix D, translates this proximity to conditional entropies.

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 $0 \le \varepsilon < \frac{1}{6}$,*

$$(1-\varepsilon)q(a) \le p(a) \le (1+\varepsilon)q(a), \quad \forall a \in \mathcal{A}. \tag{68}$$

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

$$\left| H(f(A)|g(A)) - H(f(\tilde{A})|g(\tilde{A})) \right| < \sqrt{8\varepsilon}.$$

We are now ready to prove Theorem 18.

Proof of Theorem 18: Choose $\varepsilon_1 > 0$ and $0 < \varepsilon_2 < \frac{1}{6}$ small enough such that

$$\xi \triangleq \eta - 4\varepsilon_1 - (2\varepsilon_1 + \sqrt{8\varepsilon_2}) > 0. \tag{69}$$

For example, one may take $\varepsilon_1 < \eta/12$ and $\varepsilon_2 < \eta^2/32$. Take n_{th} large enough so that Proposition 15 holds with ξ 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}, ε_1, and ε_2 above, there exist L and M such that (62) holds for $L_0 = L$ and $M_0 = M$. That is,

$$-(2\varepsilon_1 + \sqrt{8\varepsilon_2}) \le H(F_i|G_i) - H(\tilde{F}_i|\tilde{G}_i) \le (2\varepsilon_1 + \sqrt{8\varepsilon_2}). \tag{70}$$

In fact, we choose $L_0 = L$ as in the proof of Lemma 20. This ensures that the L_0-forgetfulness of the s/o-process is upper-bounded by ε_1. Thus, by Corollary 10, (41) holds with $\epsilon \le \varepsilon_1$, so that

$$-2\varepsilon_1 \le \mathcal{H}(X_\star|Y_\star) - \tilde{\mathcal{H}} \le 2\varepsilon_1.$$

Hence, if $\mathcal{H}(X_\star|Y_\star) \le 1/2$ then $\tilde{\mathcal{H}} \le (1+4\varepsilon_1)/2$ and if $\mathcal{H}(X_\star|Y_\star) \ge 1/2$ then $\tilde{\mathcal{H}} \ge (1-4\varepsilon_1)/2$. Consequently, by Corollary 16 with $\zeta = 4\varepsilon_1$, if $n \ge n_{\text{th}}$ then

$$\begin{aligned} \mathcal{H}(X_\star|Y_\star) \le 1/2 &\Rightarrow H(\tilde{F}_i|\tilde{G}_i) < \xi + 4\varepsilon_1, & \forall i \in [\text{med}_+(n)], \\ \mathcal{H}(X_\star|Y_\star) \ge 1/2 &\Rightarrow H(\tilde{F}_i|\tilde{G}_i) > 1 - \xi - 4\varepsilon_1, & \forall i \in [\text{med}_-(n)]. \end{aligned}$$

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

$$\begin{aligned} \mathcal{H}(X_\star|Y_\star) \le 1/2 &\Rightarrow H(F_i|G_i) < \eta, & \forall i \in [\text{med}_+(n)], \\ \mathcal{H}(X_\star|Y_\star) \ge 1/2 &\Rightarrow H(F_i|G_i) > 1 - \eta, & \forall i \in [\text{med}_-(n)]. \end{aligned}$$

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 $|\mathcal{S}|$ states and the code length is $\hat{N}$, the decoding complexity is $O(|\mathcal{S}|^3\hat{N}\cdot\log\hat{N})$, 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 $O(|\mathcal{S}|^3 N\hat{N}\cdot\log(N\hat{N}))$. As mentioned in Section III-C, the overall decoding error of this scheme is upper-bounded by $N\hat{N}\cdot 2^{-\hat{N}^\beta}$ 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 $\tau(\mathsf{M}) < 1$ such that $d(\mathbf{x}^T\mathsf{M}, \mathbf{y}^T\mathsf{M}) \le \tau(\mathsf{M})d(\mathbf{x},\mathbf{y})$.

This section invariably contains a large number of indices. For tractability, we adhere to the following notational convention 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* $\sigma(\mathbf{x})$ of a vector $\mathbf{x}$ is the set of its nonzero indices. That is, $\sigma(\mathbf{x}) = \{i \mid x_i \neq 0\}$. 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 $\mathbf{x}, \mathbf{y}$ be two nonnegative nonzero vectors such that $\sigma(\mathbf{x}) = \sigma(\mathbf{y})$. The *projective distance* d between the two vectors is

$$d(\mathbf{x},\mathbf{y}) \triangleq \max_{j,l\in\sigma(\mathbf{x})} \ln\frac{x_j/y_j}{x_l/y_l} = \ln \max_{j,l\in\sigma(\mathbf{x})} \frac{x_j/y_j}{x_l/y_l}. \tag{71}$$

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

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 $\mathbf{x}$ and $\mathbf{y}$ 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 $d(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = c\mathbf{y}$ 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.

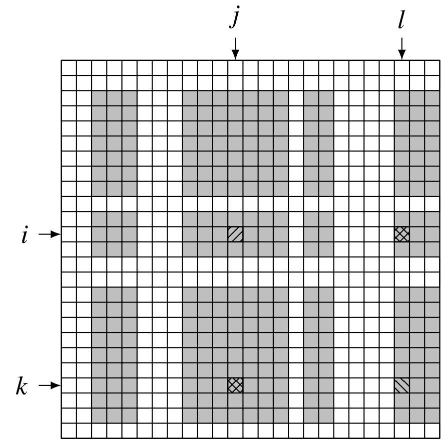

Fig. 10. An illustration of a subrectangular matrix. Each of the small squares is an element of the matrix. The white squares contain zeros, whereas the filled squares contain positive values. Elements $(\mathsf{M})_{i,j}$ and $(\mathsf{M})_{k,l}$, denoted with diagonal lines (▨ and ▧ respectively), are nonzero. Therefore, elements $(\mathsf{M})_{i,l}$ and $(\mathsf{M})_{k,j}$, denoted with a crosshatch (▩), are also nonzero. In fact, any matrix element in the support of a subrectangular matrix is nonzero.

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

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 $\mathcal{N}_\mathrm{r}(\mathsf{M})$ and its set of nonzero columns by $\mathcal{N}_\mathrm{c}(\mathsf{M})$.

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:*

$$(\mathsf{M})_{i,j} = 0 \iff i \notin \mathcal{N}_\mathrm{r}(\mathsf{M}) \text{ or } j \notin \mathcal{N}_\mathrm{c}(\mathsf{M}). \tag{72}$$

3) *The matrix M satisfies*

$$(\mathsf{M})_{i,j} \neq 0 \iff i \in \mathcal{N}_\mathrm{r}(\mathsf{M}) \text{ and } j \in \mathcal{N}_\mathrm{c}(\mathsf{M}). \tag{73}$$

Proof: The second and third characterizations are clearly equivalent. Hence, it suffices to show that $1 \Rightarrow 2$ and $3 \Rightarrow 1$.

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

$3 \Rightarrow 1$: 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 $(\mathsf{M})_{i,j} \neq 0$ and $(\mathsf{M})_{k,l} \neq 0$. Thus, by (73), $i,k \in \mathcal{N}_\mathrm{r}(\mathsf{M})$ and $j,l \in \mathcal{N}_\mathrm{c}(\mathsf{M})$. Then, a second of use of (73) implies that $(\mathsf{M})_{i,l} \neq 0$ and $(\mathsf{M})_{k,j} \neq 0$. 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 $\mathbf{x}, \mathbf{y}$ are nonnegative vectors such that $\|\mathbf{x}^T\mathsf{M}\|_1 > 0$ and $\|\mathbf{y}^T\mathsf{M}\|_1 > 0$, then $\sigma(\mathbf{x}^T\mathsf{M}) = \sigma(\mathbf{y}^T\mathsf{M})$ and $\sigma(\mathsf{M}\mathbf{x}) = \sigma(\mathsf{M}\mathbf{y})$.*

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

Proof: It suffices to prove the claim that $\sigma(\mathbf{x}^T\mathsf{M}) = \sigma(\mathbf{y}^T\mathsf{M})$, 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 $\mathbf{x}$ and $\mathbf{y}$ 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 nonnegative and nonzero vector $\mathbf{z}$, we have $(\mathbf{z}^T\mathsf{M})_i = 0$ if and only if column i of M is an all-zero column. The claim follows since both $\mathbf{x}$ and $\mathbf{y}$ are nonnegative and nonzero. ■

The following corollary was stated as [19, 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 MT' are subrectangular.*

Proof: The case where either matrix is the zero 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 $\mathsf{T}_1, \mathsf{T}_2$ below are not subrectangular whereas their product $\mathsf{T}_1\mathsf{T}_2$ is:

$$\mathsf{T}_1 = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix}, \quad \mathsf{T}_2 = \begin{bmatrix} * & * \\ 0 & * \end{bmatrix}, \quad \mathsf{T}_1\mathsf{T}_2 = \begin{bmatrix} * & * \\ * & * \end{bmatrix}.$$

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

$$\sigma(\mathsf{M}) = \{(i, j) \mid i \in \mathcal{N}_{\mathrm{r}}(\mathsf{M}), j \in \mathcal{N}_{\mathrm{c}}(\mathsf{M})\}.$$

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

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

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

By Lemma 24 and the positivity of $\mathbf{x}$ and $\mathbf{y}$, the numerator of (74) is well-defined. That is, $\mathbf{x}^T\mathsf{M}$ and $\mathbf{y}^T\mathsf{M}$ have the same support. The denominator of (74) is also well-defined, as $\mathbf{x}$ and $\mathbf{y}$ 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 $\mathbf{x} \neq c\mathbf{y}$ for $c > 0$.

The Birkhoff contraction coefficient [34, 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

$$\tau(\mathsf{M}) = \tau(\mathsf{M}^T), \tag{75}$$

since $d(\mathbf{x}^T\mathsf{M}, \mathbf{y}^T\mathsf{M}) = d(\mathsf{M}^T\mathbf{x}, \mathsf{M}^T\mathbf{y})$.

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

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

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

where

$$\phi(\mathsf{M}) \triangleq \min_{\substack{i,k \in \mathcal{N}_{\mathrm{r}}(\mathsf{M}),\\ j,l \in \mathcal{N}_{\mathrm{c}}(\mathsf{M})}} \frac{(\mathsf{M})_{i,j}(\mathsf{M})_{k,l}}{(\mathsf{M})_{i,l}(\mathsf{M})_{k,j}} > 0. \tag{76}$$

Since M is subrectangular and nonzero, all index pairs on the right-hand side of (76) are in the support of M, by which $\phi(\mathsf{M}) > 0$. 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, Section 3.4] and [36, 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 $\mathbf{x}$ and $\mathbf{y}$ are *positive* vectors and M is subrectangular, then

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

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

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

$$d(\mathbf{x}^T\mathsf{M}, \mathbf{y}^T\mathsf{M}) \leq \tau(\mathsf{M})d(\mathbf{x}, \mathbf{y}). \tag{77}$$

Proof: The claim is trivial if $\mathbf{x} = \mathbf{y} = \mathbf{0}$. If $\mathbf{x}, \mathbf{y}$ are positive, the claim follows from Definition 14 and Theorem 26. So, we assume that $\mathbf{x}$ and $\mathbf{y}$ are nonzero but have some zero elements. Denote by $\tilde{\mathbf{x}}, \tilde{\mathbf{y}}$ the vectors formed from $\mathbf{x}, \mathbf{y}$ by deleting their zero elements, and by $\tilde{\mathsf{M}}$ 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,

$$\begin{aligned} d(\mathbf{x}^T\mathsf{M}, \mathbf{y}^T\mathsf{M}) &= d(\tilde{\mathbf{x}}^T\tilde{\mathsf{M}}, \tilde{\mathbf{y}}^T\tilde{\mathsf{M}}) \\ &\le \tau(\tilde{\mathsf{M}})d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \tau(\tilde{\mathsf{M}})d(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Finally, observe that $(1-\sqrt{x})/(1+\sqrt{x})$ is a decreasing function of x when $x \ge 0$; this is easily seen by computing its derivative, $-(\sqrt{x}(1+\sqrt{x})^2)^{-1}$. Since $\tilde{\mathsf{M}}$ is formed from M by deleting rows, $\phi(\tilde{\mathsf{M}}) \ge \phi(\mathsf{M})$. Thus, we must have $\tau(\tilde{\mathsf{M}}) \le \tau(\mathsf{M})$, 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 $\gamma_i \ge 0$ for all i. Assume that $\gamma_i > 0$ for some i. Then,*

$$\min_i \frac{\alpha_i}{\beta_i} \le \frac{\sum_i \gamma_i\alpha_i}{\sum_i \gamma_i\beta_i} \le \max_i \frac{\alpha_i}{\beta_i}. \tag{78}$$

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

$$\frac{\sum_i \gamma_i\alpha_i}{\sum_i \gamma_i\beta_i} = \frac{\sum_i \gamma_i\beta_i\rho_i}{\sum_i \gamma_i\beta_i} = \sum_i \frac{\gamma_i\beta_i}{\sum_{i'} \gamma_{i'}\beta_{i'}}\rho_i = \sum_i \theta_i\rho_i,$$

where $\theta_i \ge 0$ 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 ρ_i. Hence, it is lower-bounded by $\min_i \rho_i$ and upper-bounded by $\max_i \rho_i$, as required. ∎

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,*

$$\tau(\mathsf{TM}) \le \tau(\mathsf{M}).$$

If, in addition, T is subrectangular then

$$\tau(\mathsf{TM}) \le \tau(\mathsf{T})\tau(\mathsf{M}). \tag{79}$$

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), $\tau(\mathsf{TM}) = \tau((\mathsf{TM})^T) = \tau(\mathsf{M}^T\mathsf{T}^T)$ and M is subrectangular if and only if M^T is subrectangular.

Proof: There is nothing to prove if $\mathsf{TM} = 0$, so we assume that TM is nonzero.

By Corollary 25, TM is subrectangular. For the first claim, let $i_0, k_0 \in \mathcal{N}_\mathrm{r}(\mathsf{TM})$ and $j_0, l_0 \in \mathcal{N}_\mathrm{c}(\mathsf{TM})$ achieve the minimum in (76); that is, be such that $\phi(\mathsf{TM}) = ((\mathsf{TM})_{i_0,j_0}(\mathsf{TM})_{k_0,l_0})/((\mathsf{TM})_{i_0,l_0}(\mathsf{TM})_{k_0,j_0})$. Thus, by (73), $(i_0, j_0), (k_0, l_0) \in \sigma(\mathsf{TM})$. This implies that $j_0, l_0 \in \mathcal{N}_\mathrm{c}(\mathsf{M})$ — otherwise, for example, we would have $(\mathsf{TM})_{i_0,j_0} = \sum_r (\mathsf{T})_{i_0,r}(\mathsf{M})_{r,j_0} = 0$, which contradicts $(i_0, j_0) \in \sigma(\mathsf{TM})$.

Hence,

$$\begin{aligned} \phi(\mathsf{TM}) &= \frac{(\mathsf{TM})_{i_0,j_0}(\mathsf{TM})_{k_0,l_0}}{(\mathsf{TM})_{i_0,l_0}(\mathsf{TM})_{k_0,j_0}} \\ &= \frac{\sum_i (\mathsf{T})_{i_0,i}(\mathsf{M})_{i,j_0}}{\sum_i (\mathsf{T})_{i_0,i}(\mathsf{M})_{i,l_0}} \cdot \frac{\sum_k (\mathsf{T})_{k_0,k}(\mathsf{M})_{k,l_0}}{\sum_k (\mathsf{T})_{k_0,k}(\mathsf{M})_{k,j_0}} \\ &= \frac{\sum_{i\in\mathcal{N}_\mathrm{r}(\mathsf{M})} (\mathsf{T})_{i_0,i}(\mathsf{M})_{i,j_0}}{\sum_{i\in\mathcal{N}_\mathrm{r}(\mathsf{M})} (\mathsf{T})_{i_0,i}(\mathsf{M})_{i,l_0}} \cdot \frac{\sum_{k\in\mathcal{N}_\mathrm{r}(\mathsf{M})} (\mathsf{T})_{k_0,k}(\mathsf{M})_{k,l_0}}{\sum_{k\in\mathcal{N}_\mathrm{r}(\mathsf{M})} (\mathsf{T})_{k_0,k}(\mathsf{M})_{k,j_0}} \\ &\overset{\text{(a)}}{\ge} \min_{i,k\in\mathcal{N}_\mathrm{r}(\mathsf{M})} \frac{(\mathsf{M})_{i,j_0}(\mathsf{M})_{k,l_0}}{(\mathsf{M})_{i,l_0}(\mathsf{M})_{k,j_0}} \\ &\overset{\text{(b)}}{\ge} \min_{\substack{i,k\in\mathcal{N}_\mathrm{r}(\mathsf{M}) \\ j,l\in\mathcal{N}_\mathrm{c}(\mathsf{M})}} \frac{(\mathsf{M})_{i,j}(\mathsf{M})_{k,l}}{(\mathsf{M})_{i,l}(\mathsf{M})_{k,j}} \\ &= \phi(\mathsf{M}), \end{aligned}$$

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

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

$$\begin{aligned} d(\mathbf{x}^T\mathsf{TM}, \mathbf{y}^T\mathsf{TM}) &= d((\mathbf{x}^T\mathsf{T})\mathsf{M}, (\mathbf{y}^T\mathsf{T})\mathsf{M}) \\ &\le \tau(\mathsf{M})d(\mathbf{x}^T\mathsf{T}, \mathbf{y}^T\mathsf{T}) \\ &\le \tau(\mathsf{M})\tau(\mathsf{T})d(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Thus, by (74), $\tau(\mathsf{TM}) \le \tau(\mathsf{T})\tau(\mathsf{M})$. ∎

Applying Lemma 29 to a product of m subrectangular matrices $\mathsf{M}_1, \mathsf{M}_2, \ldots, \mathsf{M}_m$, we obtain

$$\tau(\mathsf{M}_1\mathsf{M}_2\cdots\mathsf{M}_m) \le \prod_{\ell=1}^{m} \tau(\mathsf{M}_\ell). \tag{80}$$

Corollary 27 required that $\mathbf{x}, \mathbf{y}$ both have the same support. For the cases where $\mathbf{x}$ and $\mathbf{y}$ 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 $\mathbf{x}$ and $\mathbf{y}$ such that $\left\|\mathbf{x}^T\mathsf{TM}\right\|_1 > 0$ and $\left\|\mathbf{y}^T\mathsf{TM}\right\|_1 > 0$,*

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

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

Proof: There is nothing to prove if $\mathsf{TM} = 0$, so we assume that TM is nonzero. By Corollary 25, $\tilde{\mathsf{M}} = \mathsf{TM}$ is subrectangular.

Fix any $i_0 \in \mathcal{N}_\mathrm{r}(\tilde{\mathsf{M}})$. Such an i_0 must exist because $\tilde{\mathsf{M}}$ is subrectangular and $\mathbf{x}^T\tilde{\mathsf{M}}$ is nonzero by assumption. By Lemma 24, and subrectangularity of $\tilde{\mathsf{M}}$,

$$\sigma(\mathbf{e}_{i_0}^T\tilde{\mathsf{M}}) = \sigma(\mathbf{x}^T\tilde{\mathsf{M}}) = \mathcal{N}_\mathrm{c}(\tilde{\mathsf{M}}). \tag{82}$$

By the symmetry and triangle inequality properties of the projective distance [34, Exercise 3.1],

$$d(\mathbf{x}^T\tilde{\mathsf{M}}, \mathbf{y}^T\tilde{\mathsf{M}}) \le d(\mathbf{e}_{i_0}^T\tilde{\mathsf{M}}, \mathbf{x}^T\tilde{\mathsf{M}}) + d(\mathbf{e}_{i_0}^T\tilde{\mathsf{M}}, \mathbf{y}^T\tilde{\mathsf{M}}).$$

Thus, by Lemma 29 and since $\ln((1+x)/(1-x))$ is monotone increasing for $0 \le x < 1$, (81) will follow if we show that

$$d(\mathbf{e}_{i_0}^T\tilde{\mathsf{M}}, \mathbf{x}^T\tilde{\mathsf{M}}) \le \ln\left(\frac{1}{\phi(\tilde{\mathsf{M}})}\right) = 2\ln\left(\frac{1+\tau(\tilde{\mathsf{M}})}{1-\tau(\tilde{\mathsf{M}})}\right),$$

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

For any $j \in \mathcal{N}_c(\tilde{\mathsf{M}})$ denote

$$\rho_j = \frac{(\mathbf{e}_{i_0}^T\tilde{\mathsf{M}})_j}{(\mathbf{x}^T\tilde{\mathsf{M}})_j} = \frac{(\tilde{\mathsf{M}})_{i_0,j}}{\sum\limits_{k\in\mathcal{N}_r(\tilde{\mathsf{M}})} x_k(\tilde{\mathsf{M}})_{k,j}}.$$

The denominator is positive by (82), so ρ_j is well-defined. Now, for $j, l \in \mathcal{N}_c(\tilde{\mathsf{M}})$,

$$\begin{aligned}\frac{\rho_j}{\rho_l} &= \frac{\sum\limits_{k\in\mathcal{N}_r(\tilde{\mathsf{M}})} x_k(\tilde{\mathsf{M}})_{k,l}}{\sum\limits_{k\in\mathcal{N}_r(\tilde{\mathsf{M}})} x_k(\tilde{\mathsf{M}})_{k,j}} \cdot \frac{(\tilde{\mathsf{M}})_{i_0,j}}{(\tilde{\mathsf{M}})_{i_0,l}} \\ &\overset{(a)}{\le} \max_{k\in\mathcal{N}_r(\tilde{\mathsf{M}})} \frac{(\tilde{\mathsf{M}})_{k,l}}{(\tilde{\mathsf{M}})_{k,j}} \cdot \frac{(\tilde{\mathsf{M}})_{i_0,j}}{(\tilde{\mathsf{M}})_{i_0,l}} \\ &\overset{(b)}{\le} \max_{k\in\mathcal{N}_r(\tilde{\mathsf{M}})} \frac{(\tilde{\mathsf{M}})_{k,l}}{(\tilde{\mathsf{M}})_{k,j}} \cdot \max_{i\in\mathcal{N}_r(\tilde{\mathsf{M}})} \frac{(\tilde{\mathsf{M}})_{i,j}}{(\tilde{\mathsf{M}})_{i,l}}, \end{aligned} \tag{83}$$

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),

$$\begin{aligned} d(\mathbf{e}_{i_0}^T\tilde{\mathsf{M}}, \mathbf{x}^T\tilde{\mathsf{M}}) &= \ln \max_{j,l\in\mathcal{N}_c(\tilde{\mathsf{M}})} \frac{\rho_j}{\rho_l} \\ &\overset{(a)}{\le} \ln \max_{\substack{i,k\in\mathcal{N}_r(\tilde{\mathsf{M}}),\\ j,l\in\mathcal{N}_c(\tilde{\mathsf{M}})}} \frac{(\tilde{\mathsf{M}})_{i,j}(\tilde{\mathsf{M}})_{k,l}}{(\tilde{\mathsf{M}})_{i,l}(\tilde{\mathsf{M}})_{k,j}} \\ &\overset{(b)}{=} \ln\left(\frac{1}{\phi(\tilde{\mathsf{M}})}\right), \end{aligned}$$

where (a) is by (83) and (b) follows from the definition of ϕ in (76). This completes the proof. ∎

The following proposition and the corollary that follows are a generalization of ideas from [18, Theorem 2].

Proposition 31. *Let $\mathsf{M}_1, \mathsf{M}_2, \ldots, \mathsf{M}_m, \mathsf{T}$ be a sequence of square nonzero nonnegative matrices, such that M_ℓ are subrectangular for all $1 \le \ell \le m$, and let $\mathbf{x}, \mathbf{y}$ be two nonnegative nonzero vectors. Denote*

$$\begin{aligned}\tilde{\mathbf{x}}^T &= \mathbf{x}^T\mathsf{M}_1, \\ \tilde{\mathbf{y}}^T &= \mathbf{y}^T\mathsf{M}_1, \\ \mathsf{M}_r^s &= \mathsf{M}_r \cdot \mathsf{M}_{r+1}\cdots\mathsf{M}_s, \quad r \le s.\end{aligned}$$

If $\left\|\mathbf{x}^T\mathsf{M}_1^m\mathsf{T}\right\|_1 > 0$ and $\left\|\mathbf{y}^T\mathsf{M}_1^m\mathsf{T}\right\|_1 > 0$, then

$$\ln\left(\frac{\left\|\mathbf{x}^T\mathsf{M}_1^m\mathsf{T}\right\|_1}{\left\|\mathbf{y}^T\mathsf{M}_1^m\mathsf{T}\right\|_1} \cdot \frac{\left\|\mathbf{y}^T\mathsf{M}_1^m\right\|_1}{\left\|\mathbf{x}^T\mathsf{M}_1^m\right\|_1}\right) \le d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \cdot \prod_{\ell=2}^{m} \tau(\mathsf{M}_\ell). \tag{84}$$

Proof: Since $\left\|\mathbf{x}^T\mathsf{M}_1^m\mathsf{T}\right\|_1 > 0$, we conclude that $\mathbf{x}^T\mathsf{M}_1^s$ is nonzero for any $1 \le s \le m$, and the same holds if we replace $\mathbf{x}$ with $\mathbf{y}$. Thus, the left-hand side of (84) is well-defined. We will show that

$$\ln\left(\frac{\left\|\mathbf{x}^T\mathsf{M}_1^m\mathsf{T}\right\|_1}{\left\|\mathbf{y}^T\mathsf{M}_1^m\mathsf{T}\right\|_1} \cdot \frac{\left\|\mathbf{y}^T\mathsf{M}_1^m\right\|_1}{\left\|\mathbf{x}^T\mathsf{M}_1^m\right\|_1}\right) \le d(\tilde{\mathbf{x}}^T\mathsf{M}_2^m, \tilde{\mathbf{y}}^T\mathsf{M}_2^m).$$

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

Denote $J = \sigma(\tilde{\mathbf{x}}^T\mathsf{M}_2^m) = \sigma(\tilde{\mathbf{y}}^T\mathsf{M}_2^m) = \mathcal{N}_c(\mathsf{M}_2^m)$, where the equalities are by Lemma 24 and subrectangularity. By the right-hand inequality of (78),

$$\begin{aligned}\frac{\left\|\mathbf{y}^T\mathsf{M}_1^m\right\|_1}{\left\|\mathbf{x}^T\mathsf{M}_1^m\right\|_1} &= \frac{\left\|\tilde{\mathbf{y}}^T\mathsf{M}_2^m\right\|_1}{\left\|\tilde{\mathbf{x}}^T\mathsf{M}_2^m\right\|_1} \\ &= \frac{\sum_{l\in J} 1\cdot(\tilde{\mathbf{y}}^T\mathsf{M}_2^m)_l}{\sum_{l\in J} 1\cdot(\tilde{\mathbf{x}}^T\mathsf{M}_2^m)_l} \le \max_{l\in J}\frac{(\tilde{\mathbf{y}}^T\mathsf{M}_2^m)_l}{(\tilde{\mathbf{x}}^T\mathsf{M}_2^m)_l}.\end{aligned}$$

Next, denote by $t_j = \left\|(\mathsf{T})_{j,:}\right\|_1$ the sum of the jth 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

$$\frac{\left\|\mathbf{x}^T\mathsf{M}_1^m\mathsf{T}\right\|_1}{\left\|\mathbf{y}^T\mathsf{M}_1^m\mathsf{T}\right\|_1} = \frac{\sum_{j\in J} t_j\cdot(\tilde{\mathbf{x}}^T\mathsf{M}_2^m)_j}{\sum_{j\in J} t_j\cdot(\tilde{\mathbf{y}}^T\mathsf{M}_2^m)_j} \le \max_{j\in J}\frac{(\tilde{\mathbf{x}}^T\mathsf{M}_2^m)_j}{(\tilde{\mathbf{y}}^T\mathsf{M}_2^m)_j}.$$

Combining, we obtain

$$\frac{\left\|\mathbf{x}^T\mathsf{M}_1^m\mathsf{T}\right\|_1}{\left\|\mathbf{y}^T\mathsf{M}_1^m\mathsf{T}\right\|_1} \cdot \frac{\left\|\mathbf{y}^T\mathsf{M}_1^m\right\|_1}{\left\|\mathbf{x}^T\mathsf{M}_1^m\right\|_1} \le \max_{j,l\in J}\frac{(\tilde{\mathbf{x}}^T\mathsf{M}_2^m)_j/(\tilde{\mathbf{y}}^T\mathsf{M}_2^m)_j}{(\tilde{\mathbf{x}}^T\mathsf{M}_2^m)_l/(\tilde{\mathbf{y}}^T\mathsf{M}_2^m)_l}.$$

Taking the logarithm of both sides, the right-hand side becomes $d(\tilde{\mathbf{x}}^T\mathsf{M}_2^m, \tilde{\mathbf{y}}^T\mathsf{M}_2^m)$, which completes the proof. ∎

Combining the above results we obtain the following corollary.

Corollary 32. *Let $\mathsf{M}_1, \mathsf{M}_2, \ldots, \mathsf{M}_m$ be a sequence of square nonzero subrectangular matrices, and let T, as well as $\mathsf{T}_1, \mathsf{T}_2, \ldots, \mathsf{T}_m$ be arbitrary square nonnegative and nonzero matrices. Denote*

$$\mathsf{R} = \mathsf{T}_1\mathsf{M}_1\mathsf{T}_2\mathsf{M}_2\cdots\mathsf{T}_m\mathsf{M}_m.$$

Then, for any two nonnegative nonzero vectors $\mathbf{x}, \mathbf{y}$ such that $\left\|\mathbf{x}^T\mathsf{R}\mathsf{T}\right\|_1 > 0$ and $\left\|\mathbf{y}^T\mathsf{R}\mathsf{T}\right\|_1 > 0$ we have

$$\ln\left(\frac{\left\|\mathbf{x}^T\mathsf{R}\mathsf{T}\right\|_1}{\left\|\mathbf{y}^T\mathsf{R}\mathsf{T}\right\|_1} \cdot \frac{\left\|\mathbf{y}^T\mathsf{R}\right\|_1}{\left\|\mathbf{x}^T\mathsf{R}\right\|_1}\right) \le 4\ln\left(\frac{1+\tau(\mathsf{M}_1)}{1-\tau(\mathsf{M}_1)}\right) \cdot \prod_{\ell=2}^{m} \tau(\mathsf{M}_\ell). \tag{85}$$

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 32 generalize [18, Theorem 2] in several ways. In [18, Theorem 2], the matrices $\mathsf{M}_1, \ldots, \mathsf{M}_m, \mathsf{T}$ 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

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

This is a stochastic matrix: $(\mathsf{M})_{i,j} \geq 0$ for all $i, j \in \mathcal{A}$ and for any i, $\sum_j (\mathsf{M})_{i,j} = 1$. 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 $\boldsymbol{\pi}$, 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 $B \subseteq \mathcal{B}$,

$$f^{-1}(B) = \{i \in \mathcal{A} \mid f(i) = b,\ b \in B\}.$$

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

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

[7] 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.

Definition 15 (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.

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 $\mathsf{M}(b)$, $b \in \mathcal{B}$, by

$$(\mathsf{M}(b))_{i,j} = \begin{cases} p(j|i), & \text{if } f(j) = b \\ 0, & \text{otherwise.} \end{cases} \tag{86}$$

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

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

For a sequence of observations b_r^s, $r \leq s$, we denote

$$\mathsf{M}(b_r^s) \triangleq \mathsf{M}(b_r)\mathsf{M}(b_{r+1}) \cdots \mathsf{M}(b_s).$$

We call $\tau(\mathsf{M}(b_r^s))$ the Birkhoff contraction coefficient *induced* by the sequence b_r^s.

The matrices $\mathsf{M}(b)$ 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

$$\mathsf{M} = \sum_b \mathsf{M}(b).$$

Example 8 in appendix E shows the matrix M and its decomposition to matrices $\mathsf{M}(b)$ 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 $\mathsf{M}(b)$, $b \in \mathcal{B}$, and its initial distribution. We assume that the process is stationary, so its initial distribution is $\boldsymbol{\pi}$, its unique stationary distribution. Thus, $(\boldsymbol{\pi})_i = \mathbb{P}(A_0 = i)$ and

$$\begin{aligned}\mathbb{P}(B_1 = b_1) &= \sum_{j\in\mathcal{A}} \mathbb{P}(A_1 = j, B_1 = b_1)\\ &= \sum_{i,j\in\mathcal{A}} \mathbb{P}(A_1 = j, B_1 = b_1 | A_0 = i)\,\mathbb{P}(A_0 = i)\\ &= \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1)\right\|_1\end{aligned}$$

Moreover, the probability of observing the sequence b_1^n is given by [19, Lemma 2.1]

$$\begin{aligned}\mathbb{P}\left(B_1^n = b_1^n\right) &= \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\right\|_1\\ &= \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1)\mathsf{M}(b_2)\cdots\mathsf{M}(b_n)\right\|_1. \end{aligned} \tag{87}$$

Similarly, for any $a \in \mathcal{A}$,

$$\mathbb{P}\left(A_n = a, B_1^n = b_1^n\right) = (\boldsymbol{\pi}^T \mathsf{M}(b_1^n))_a = \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\mathbb{I}_a\right\|_1,$$

and

$$\begin{aligned}\mathbb{P}\left(A_{n+1} = a, B_1^n = b_1^n\right) &= (\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\mathsf{M})_a\\ &= \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\mathsf{M}\mathbb{I}_a\right\|_1\\ &= \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\mathsf{T}_a\right\|_1, \end{aligned} \tag{88}$$

where we denoted for any $a \in \mathcal{A}$,

$$\mathsf{T}_a \triangleq \mathsf{M}\mathbb{I}_a.$$

When $\mathbb{P}\left(B_1^n = b_1^n\right) > 0$ we further have by (87) and (88),

$$\begin{aligned}\mathbb{P}\left(A_{n+1} = a | B_1^n = b_1^n\right) &= \frac{\mathbb{P}\left(A_{n+1} = a, B_1^n = b_1^n\right)}{\mathbb{P}\left(B_1^n = b_1^n\right)}\\ &= \frac{\left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\mathsf{T}_a\right\|_1}{\left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\right\|_1}. \end{aligned} \tag{89}$$

This is well-defined because if $\mathbb{P}\left(B_1^n = b_1^n\right) > 0$ 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, $\mathbb{P}(B_1 = b_1 | A_0 = a_0) = \left\|\mathbf{e}_{a_0}^T \mathsf{M}(b_1)\right\|_1$. Similar to the above, we obtain

$$\mathbb{P}\left(B_1^n = b_1^n | A_0 = a_0\right) = \left\|\mathbf{e}_{a_0}^T \mathsf{M}(b_1^n)\right\|_1, \tag{90}$$

$$\mathbb{P}\left(A_{n+1} = a | B_1^n = b_1^n, A_0 = a_0\right) = \frac{\left\|\mathbf{e}_{a_0}^T \mathsf{M}(b_1^n)\mathsf{T}_a\right\|_1}{\left\|\mathbf{e}_{a_0}^T \mathsf{M}(b_1^n)\right\|_1}, \tag{91}$$

provided that the probability in (90) is positive.

In (87)–(91), we have computed probabilities for particular realizations of A_0, B_1^n and A_{n+1}. Generally, however, these are random variables. They are jointly generated as follows. First, draw A_0 according to $\boldsymbol{\pi}$. 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 $\mathbb{P}\left(A_{n+1} | B_1^n\right)$ and $\mathbb{P}\left(A_{n+1} | B_1^n, A_0\right)$, obtained by plugging A_{n+1}, B_1^n, and A_0 for a, b_1^n, 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

$$I(A_0; A_{n+1} | B_1^n) = \mathbb{E}\left[\log \frac{\mathbb{P}\left(A_{n+1} | B_1^n, A_0\right)}{\mathbb{P}\left(A_{n+1} | B_1^n\right)}\right]. \tag{92}$$

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

$$\begin{aligned}&I(A_0; A_{n+1} | B_1^n)\\ &= \mathbb{E}\left[\log\left(\frac{\left\|\mathbf{e}_{A_0}^T \mathsf{M}(B_1^n)\mathsf{T}_{A_{n+1}}\right\|_1}{\left\|\boldsymbol{\pi}^T \mathsf{M}(B_1^n)\mathsf{T}_{A_{n+1}}\right\|_1} \cdot \frac{\left\|\boldsymbol{\pi}^T \mathsf{M}(B_1^n)\right\|_1}{\left\|\mathbf{e}_{A_0}^T \mathsf{M}(B_1^n)\right\|_1}\right)\right]. \end{aligned} \tag{93}$$

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 $I(A_0; A_{n+1}) \to 0$ as $n \to \infty$. By the Markov property we also have $I(A_0; A_{n+1} | A_1^n) = 0$. Our setting, however, is a hidden Markov setting, and we will be interested in whether $I(A_0; A_{n+1} | B_1^n) \to 0$. 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, $I(A_0; A_{n+1} | B_1^n) \to 0$ as $n \to \infty$. 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 $I(A_0; B_n | B_1^{n-1}) \to 0$. 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 $\mathsf{M}(b)$, $b \in \mathcal{B}$, contain only two types of columns: strictly positive columns and zero columns.[9] In this case, the matrices $\mathsf{M}(b)$ are all subrectangular, so their Birkhoff contraction coefficients are strictly less than 1; this allows one to use (85) directly (with $\mathsf{T}_\ell = \mathbb{I}$ for all ℓ) 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 $\mathsf{M}(b)$ may also be *not* subrectangular. We further remark that, by the data processing inequality (2), $I(A_0; A_{n+1} | B_1^n) \to 0$ implies that $I(A_0; B_{n+1} | B_1^n) \to 0$.

B. Forgetting the Initial State

We now show that under the following *Condition K* (so named in honor of Prof. Thomas Kaijser who had first suggested it in [19]), the mutual information $I(A_0; A_{n+1} | B_1^n)$ vanishes with n.

[8] Where 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.

[9] The assumption of [18] is that M is positive. Since $\mathsf{M}(b)$ is comprised of columns of M and zero columns, any nonzero column of $\mathsf{M}(b)$ must be positive.

Condition K. The HMM (A_n, B_n) is characterized by matrices $\mathsf{M}(b)$, $b \in \mathcal{B}$ such that:

1) The matrix $\mathsf{M} = \sum_{b\in\mathcal{B}} \mathsf{M}(b)$ is aperiodic and irreducible.
2) There exists an ordered sequence $\beta_1, \beta_2, \ldots, \beta_l$ of elements of $\mathcal{B}$ such that the matrix $\mathsf{M}(\beta_1^l) = \mathsf{M}(\beta_1)\mathsf{M}(\beta_2)\cdots\mathsf{M}(\beta_l)$ is nonzero 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 β for which $\mathsf{M}(\beta)$ has just a single column;
- there exists an observation β for which $\mathsf{M}(\beta)$ 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 λ such that if $n \geq \lambda$ then*

$$I(A_0; A_{n+1}|B_1^n) \leq \epsilon.$$

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, $\mathsf{M}(0) = \mathsf{M}$ and Condition K is satisfied, as there exists k_0 such that $\mathsf{M}^{k_0} > 0$ [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 λ such that if $n \geq \lambda$ then*

$$I(A_1; A_n|B_1^n) \leq \epsilon. \tag{94}$$

and

$$I(A_0; A_n|B_1^n) \leq \epsilon. \tag{95}$$

Proof: The conditions of Theorem 33 are satisfied. Let λ be such that $I(A_1; A_n|B_2^{n-1}) \leq \epsilon$ 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 $n \geq \lambda$, $I(A_1, B_1; A_n, B_n|B_2^{n-1}) = I(A_1; A_n|B_2^{n-1}) \leq \epsilon$. Therefore,

$$\begin{aligned}\epsilon &\geq I(A_1, B_1; A_n, B_n|B_2^{n-1}) \\ &= I(B_1; A_n, B_n|B_2^{n-1}) + I(A_1; B_n|B_1^{n-1}) + I(A_1; A_n|B_1^n).\end{aligned}$$

Since mutual information is nonnegative, each of the summands on the right-hand side is upper-bounded by ϵ. This yields (94).

To see (95), since $A_0 -\!\!\circ\!\!- (A_1, B_1^n) -\!\!\circ\!\!- A_n$, we use (2) and obtain

$$I(A_0; A_n|B_1^n) \leq I(A_1; A_n|B_1^n) \leq \epsilon,$$

as required. ■

We remark that under the same conditions as Corollary 34 we also obtain $I(A_1; B_n|B_1^{n-1}) \leq \epsilon$ and $I(A_0; B_n|B_1^{n-1}) \leq \epsilon$.

Consider a Markov chain A_n and *two* HMMs it induces, (A_n, B_n) and (A_n, C_n), where $B_n = f(A_n)$ and $C_n = g(B_n)$, for some deterministic functions f, g. It is somewhat surprising, but even if one of the HMMs satisfies Condition K, it does not imply that the other one does. See Example 4 in Section V.[10] Suppose that both HMMs satisfy Condition K. Then, by Corollary 34, for every $\epsilon > 0$ there exists an integer λ such that if $n \geq \lambda$ then $I(A_1; A_n|B_1^n) \leq \epsilon$ and $I(A_1; A_n|C_1^n) \leq \epsilon$.

Example 7. Let (S_n, X_n, Y_n) be a FAIM process. This is an HMM with state $A_n = (S_n, X_n, Y_n)$. Clearly, there exist functions f, g such that $(X_n, Y_n) = f(S_n)$ and $Y_n = g(X_n, Y_n)$. Therefore, both $(A_n, (X_n, Y_n))$ and (A_n, Y_n) are HMMs. If each of the HMMs $(A_n, (X_n, Y_n))$ and (A_n, Y_n) satisfies Condition K then (94) holds with $B_n = (X_n, Y_n)$ or $B_n = Y_n$ for any n. In particular, for any $\epsilon > 0$ there exists an integer λ such that for any $k \geq \lambda$ we have

$$\begin{aligned}I(S_1; S_k|X_1^k, Y_1^k) &\leq \epsilon, \\ I(S_1; S_k|Y_1^k) &\leq \epsilon.\end{aligned}$$

In other words, Condition K is a sufficient condition for forgetfulness.

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 ($(n_\star, \delta_\star, \tau_\star)$-KHMM)**.** Let $n_\star$ be a positive integer, and $\delta_\star, \tau_\star \in [0, 1)$. The HMM (A_n, B_n) is called an $(n_\star, \delta_\star, \tau_\star)$-KHMM if it satisfies

$$\mathbb{P}\left(\tau(\mathsf{M}(B_1^{n_\star})) \leq \tau_\star | A_0 = a_0\right) \geq 1 - \delta_\star, \quad \forall a_0 \in \mathcal{A}. \tag{96}$$

In words, the HMM has a probability at least $(1 - \delta_\star)$ of emitting by time $n_\star$ an observation sequence that induces a Birkhoff contraction coefficient at most $\tau_\star$, regardless of its initial state.

We say that an HMM is a KHMM if it is an $(n_\star, \delta_\star, \tau_\star)$-KHMM for some $(n_\star, \delta_\star, \tau_\star)$.

[10] Taking $A_n = S_n$, $B_n = (X_n, Y_n)$, and $C_n = Y_n$.

Observe that if (A_n, B_n) is an $(n_\star, \delta_\star, \tau_\star)$-KHMM and $n_\star \le n'_\star$, $\delta_\star \le \delta'_\star$, and $\tau_\star \le \tau'_\star$ then (A_n, B_n) is also an $(n'_\star, \delta'_\star, \tau'_\star)$-KHMM.

In Lemma 35, adapted from [19, Lemma 8.2], we show that if an HMM satisfies Condition K, then it is also a KHMM for some $(n_\star, \delta_\star, \tau_\star)$. This is because Condition K ensures the existence of one 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 satifies Condition K may be an $(n_\star, \delta_\star, \tau_\star)$-KHMM for many different combinations of $n_\star, \delta_\star, \tau_\star$. Since the bounds we develop are dependent on the values of $n_\star, \delta_\star, \tau_\star$, it is worthwhile to seek the combination that yield the best bound.

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

Proof: By Condition K there exist positive integers k_0, l_0 and numbers $\gamma_0 > 0$ and $0 \le \tau_\star < 1$ such that

1) For any $i, j \in \mathcal{A}$, $(\mathsf{M}^{k_0})_{i,j} \ge \gamma_0$. This follows from M being aperiodic and irreducible, so some power of it must be strictly positive [34, Theorem 1.4].
2) For some sequence $\beta_1^{l_0}$ of elements of $\mathcal{B}$, the matrix $\mathsf{M}(\beta_1^{l_0})$ is nonzero and subrectangular. Existence of such sequences is guaranteed by Condition K. We denote $\tau_\star = \tau(\mathsf{M}(\beta_1^{l_0}))$. Since $\mathsf{M}(\beta_1^{l_0})$ is subrectangular, $0 \le \tau_\star < 1$.

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

$$\mathcal{A}' = \left\{ a \in \mathcal{A} \;\middle|\; \left\| \mathbf{e}_a^T \mathsf{M}(\beta_1^{l_0}) \right\|_1 > 0 \right\}.$$

That is, there is positive probability that the next observation after any state in $\mathcal{A}'$ is the first observation β_1 of the word $\beta_1^{l_0}$. Since Condition K is satisfied, $\mathcal{A}'$ is not empty, so that

$$\alpha_0 = \min_{a \in \mathcal{A}'} \left\| \mathbf{e}_a^T \mathsf{M}(\beta_1^{l_0}) \right\|_1 > 0.$$

We claim that (96) is satisfied with $n_\star = k_0 + l_0$ and $\delta_\star = 1 - \alpha_0\gamma_0 < 1$. Indeed, for any $a_0 \in \mathcal{A}$,

$$\begin{aligned}
&\mathbb{P}\left(\tau(\mathsf{M}(B_1^{n_\star})) \le \tau_\star \,\middle|\, A_0 = a_0\right) \\
&\overset{(a)}{\ge} \mathbb{P}\left(\tau(\mathsf{M}(B_{k_0+1}^{k_0+l_0})) \le \tau_\star \,\middle|\, A_0 = a_0\right) \\
&\overset{(b)}{\ge} \mathbb{P}\left(B_{k_0+1}^{k_0+l_0} = \beta_1^{l_0} \,\middle|\, A_0 = a_0\right) \\
&= \sum_{a \in \mathcal{A}} \mathbb{P}\left(B_{k_0+1}^{k_0+l_0} = \beta_1^{l_0}, A_{k_0} = a \,\middle|\, A_0 = a_0\right) \\
&\overset{(c)}{=} \sum_{a \in \mathcal{A}'} \mathbb{P}\left(B_{k_0+1}^{k_0+l_0} = \beta_1^{l_0} \,\middle|\, A_{k_0} = a\right) \cdot \mathbb{P}\left(A_{k_0} = a | A_0 = a_0\right) \\
&\overset{(d)}{=} \sum_{a \in \mathcal{A}'} \left\| \mathbf{e}_a^T \mathsf{M}(\beta_1^{l_0}) \right\|_1 \cdot (\mathsf{M}^{k_0})_{a_0, a} \\
&\ge \alpha_0 \gamma_0,
\end{aligned}$$

where (a) is by Corollary 25 and Lemma 29, (b) is by Condition K, (c) is by the Markov property, and (d) is by (90). ■

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

$$\begin{aligned}
N_1(\tau) &= \min\{n : \tau(\mathsf{M}(B_1^n)) \le \tau\}, \\
N_{k+1}(\tau) &= \min\{n : \tau(\mathsf{M}(B_{N_k+1}^{N_k+n})) \le \tau\}, \quad k \ge 1.
\end{aligned}$$

That is, the random variable $N_1(\tau)$ is the time of the first occurrence of a sequence that induces a Birkhoff contraction coefficient τ or less. In other words, $N_1(\tau)$ is the smallest value of n such that $\mathsf{M}(B_1^n)$ is a subrectangular matrix with Birkhoff contraction coefficient τ or less. Similarly, the random variable $N_k(\tau)$ is the gap between the $(k-1)$th and kth 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 $(n_\star, \delta_\star, \tau_\star)$-KHMM. If $\delta_\star > 0$, there exist $\gamma > 0$ and $0 \le \rho < 1$ so that for any positive integer n_1,*

$$\mathbb{P}(N_1(\tau_\star) \ge n_1 \mid A_0 = a_0) \le \gamma \rho^{n_1}, \quad \forall a_0 \in \mathcal{A}. \tag{97}$$

Proof: Let $T_0 = 1$ and denote, for any positive integer k, the random variable $T_k = \tau(\mathsf{M}(B_1^{kn_\star}))$. Observe that, by (96), $\mathbb{P}(T_1 \le \tau_\star | A_0 = a_0) \ge 1 - \delta_\star$ for any $a_0 \in \mathcal{A}$.

We now show that for any positive integer k, and any $a_0 \in \mathcal{A}$,

$$\mathbb{P}\left(T_k > \tau_\star \,\middle|\, T_{k-1} > \tau_\star, A_0 = a_0\right) \le \delta_\star. \tag{98}$$

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

$$\begin{aligned}
&\mathbb{P}\left(T_2 \le \tau_\star \,\middle|\, T_1 > \tau_\star, A_0 = a_0\right) \\
&= \mathbb{P}\left(\tau(\mathsf{M}(B_1^{2n_\star})) \le \tau_\star \,\middle|\, T_1 > \tau_\star, A_0 = a_0\right) \\
&\overset{(a)}{\ge} \mathbb{P}\left(\tau(\mathsf{M}(B_{n_\star+1}^{2n_\star})) \le \tau_\star \,\middle|\, T_1 > \tau_\star, A_0 = a_0\right) \\
&= \sum_a \mathbb{P}\left(\tau(\mathsf{M}(B_{n_\star+1}^{2n_\star})) \le \tau_\star, A_{n_\star} = a \,\middle|\, T_1 > \tau_\star, A_0 = a_0\right) \\
&\overset{(b)}{=} \sum_a \mathbb{P}\left(\tau(\mathsf{M}(B_{n_\star+1}^{2n_\star})) \le \tau_\star \,\middle|\, A_{n_\star} = a\right) p(a) \\
&\overset{(c)}{=} \sum_a \mathbb{P}\left(T_1 \le \tau_\star \mid A_0 = a\right) p(a) \\
&\overset{(d)}{\ge} 1 - \delta_\star.
\end{aligned}$$

where (a) is because, by Lemma 29, if $\tau(\mathsf{M}(B_m^n)) \le \tau_\star$ then $\tau(\mathsf{M}(B_1^n)) \le \tau_\star$; in (b) we denoted $p(a) = \mathbb{P}\left(A_{n_\star} = a | T_1 > \tau_\star, A_0 = a_0\right)$; (c) is by the Markov property; and (d) is by (96). Rearranging yields (98). We remark that (98) is also true without conditioning on $\{T_{k-1} > \tau_\star\}$.

Thus,

$$\begin{aligned}
&\mathbb{P}\left(T_k > \tau_\star | A_0 = a_0\right) \\
&= \mathbb{P}\left(T_k > \tau_\star | T_{k-1} > \tau_\star, A_0 = a_0\right) \cdot \mathbb{P}\left(T_{k-1} > \tau_\star | A_0 = a_0\right) \\
&\quad + \mathbb{P}\left(T_k > \tau_\star | T_{k-1} \le \tau_\star, A_0 = a_0\right) \cdot \mathbb{P}\left(T_{k-1} \le \tau_\star | A_0 = a_0\right) \\
&\overset{(a)}{=} \mathbb{P}\left(T_k > \tau_\star | T_{k-1} > \tau_\star, A_0 = a_0\right) \cdot \mathbb{P}\left(T_{k-1} > \tau_\star | A_0 = a_0\right) \\
&\overset{(b)}{\le} \delta_\star \mathbb{P}\left(T_{k-1} > \tau_\star | A_0 = a_0\right),
\end{aligned}$$

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 $a_0 \in \mathcal{A}$,

$$\mathbb{P}\left(N_1(\tau_\star) > kn_\star | A_0 = a_0\right) = \mathbb{P}\left(T_k > \tau_\star | A_0 = a_0\right) \le \delta_\star^k.$$

Hence, for any positive integer n_1 (not necessarily a multiple of $n_\star$) and any $a_0 \in \mathcal{A}$,

$$\mathbb{P}\left(N_1(\tau_\star) \ge n_1 | A_0 = a_0\right) \le \delta_\star^{n_1/n_\star - 1}.$$

Rearranging, this yields

$$\mathbb{P}\left(N_1(\tau_\star) \ge n_1 | A_0 = a_0\right) \le \frac{1}{\delta_\star} \cdot \left(\delta_\star^{1/n_\star}\right)^{n_1}.$$

Thus, we obtain (97) with $\gamma = 1/\delta_\star$ and $\rho = \delta_\star^{1/n_\star}$. To complete the proof, observe that $0 \le \rho < 1$ since $0 < \delta_\star < 1$. ■

We imposed $\delta_\star > 0$ in Lemma 36 because this is the more interesting case. Clearly, Lemma 36 also holds when $\delta_\star = 0$, albeit with different γ, ρ. However, we can do better in this case. Namely, if $\delta_\star = 0$ for some $n_\star$, this implies that at time $n_\star$ the sequence of observations is ensured to induce Birkhoff contraction coefficient less than $\tau_\star$. 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 $(n_\star, \delta_\star, \tau_\star)$-KHMM and $\delta_\star > 0$, we conclude that

$$\mathbb{P}\left(N_1(\tau_\star) \ge n_1\right) \le \gamma \rho^{n_1}.$$

More generally, we have the following corollary.

Corollary 37. *Let (A_n, B_n) be an $(n_\star, \delta_\star, \tau_\star)$-KHMM with $\delta_\star > 0$. Then, there exist $\gamma > 0$ and $0 \le \rho < 1$ such that for any positive integers n_1, n_2, $\ldots$, n_m,*

$$\begin{aligned}\mathbb{P}\left(N_1(\tau_\star) \ge n_1, N_2(\tau_\star) \ge n_2, \ldots, N_m(\tau_\star) \ge n_m\right) \\ \le \gamma^m \rho^{n_1+n_2+\cdots+n_m}.\end{aligned} \tag{99}$$

Proof: For brevity, we denote $N_k = N_k(\tau_\star)$. Since

$$\begin{aligned}&\mathbb{P}\left(N_1 \ge n_1, N_2 \ge n_2, \ldots, N_m \ge n_m\right) \\ &\qquad = \prod_{k=1}^{m} \mathbb{P}\left(N_k \ge n_k | N_i \ge n_i,\ i < k\right),\end{aligned}$$

(99) will follow if $\mathbb{P}\left(N_k \ge n_k | N_i \ge n_i,\ i < k\right) \le \gamma \rho^{n_k}$. Indeed, for any k we have

$$\begin{aligned}&\mathbb{P}\left(N_k \ge n_k | N_i \ge n_i,\ i < k\right) \\ &= \sum_a \mathbb{P}\left(N_k \ge n_k, A_{N_{k-1}} = a | N_i \ge n_i,\ i < k\right) \\ &= \sum_a \mathbb{P}\left(N_k \ge n_k | A_{N_{k-1}} = a\right) \mathbb{P}\left(A_{N_{k-1}} = a | N_i \ge n_i,\ i < k\right) \\ &\overset{(a)}{=} \sum_a \mathbb{P}\left(N_1 \ge n_k | A_0 = a\right) \mathbb{P}\left(A_{N_{k-1}} = a | N_i \ge n_i,\ i < k\right) \\ &\overset{(b)}{\le} \gamma \rho^{n_k} \sum_a \mathbb{P}\left(A_{N_{k-1}} = a | N_i \ge n_i,\ i < k\right) \\ &= \gamma \rho^{n_k},\end{aligned}$$

where (a) is by definition of N_k and (b) is by (97). ■

Proposition 38. *Let (A_n, B_n) be an $(n_\star, \delta_\star, \tau_\star)$-KHMM with $\delta_\star > 0$. Denote*

$$\gamma = \frac{1}{\delta_\star}, \quad \alpha = \gamma \cdot \log|\mathcal{A}|, \quad \rho = \delta_\star^{1/n_\star} < 1.$$

Then, for any $m \le n$ we have

$$I(A_0; A_{n+1} | B_1^n) \le 4 \log\left(\frac{1+\tau_\star}{1-\tau_\star}\right) \tau_\star^m + \alpha \frac{(\gamma n)^m}{m!} \rho^{n+1}. \tag{100}$$

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 $n_k \ge 1$. In [42, p. 38], it is shown that $p(n, m) = \binom{n-1}{m-1}$. Thus, by (99),

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{m} N_k(\tau_\star) \ge n\right) &\le p(n, m) \gamma^m \rho^n \\ &= \binom{n-1}{m-1} \gamma^m \rho^n \\ &\le \frac{(n-1)^{m-1}}{(m-1)!} \gamma^m \rho^n.\end{aligned}$$

Next, consider the matrix product $\mathsf{M}(B_1^n)$. 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 $\tau_\star$. This is accomplished by the integer-valued random variable

$$D_n = D_n(\tau_\star) = \max\left\{m : \sum_{k=1}^{m} N_k(\tau_\star) \le n\right\}.$$

From the above discussion,

$$\begin{aligned}\mathbb{P}\left(D_n \le m\right) &= \mathbb{P}\left(D_n < m+1\right) \\ &= \mathbb{P}\left(\sum_{k=1}^{m+1} N_k(\tau_\star) \ge n+1\right) \\ &\le \gamma \frac{(n\gamma)^m}{m!} \rho^{n+1}.\end{aligned} \tag{101}$$

Recall from (93) that $I(A_0; A_{n+1} | B_1^n) = \mathbb{E}\left[J\right]$, where we have denoted, for brevity,

$$J \triangleq \log\left(\frac{\left\|\mathbf{e}_{A_0}^T \mathsf{M}(B_1^n) \mathsf{T}_{A_{n+1}}\right\|_1}{\left\|\boldsymbol{\pi}^T \mathsf{M}(B_1^n) \mathsf{T}_{A_{n+1}}\right\|_1} \cdot \frac{\left\|\boldsymbol{\pi}^T \mathsf{M}(B_1^n)\right\|_1}{\left\|\mathbf{e}_{A_0}^T \mathsf{M}(B_1^n)\right\|_1}\right).$$

This is a conditional mutual information. In particular, for any fixed sequence b_1^n we have

$$0 \le I(A_0; A_{n+1} | B_1^n = b_1^n) = \mathbb{E}\left[J | B_1^n = b_1^n\right] \le \log|\mathcal{A}|, \tag{102}$$

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_1^n — given any realization b_1^n of B_1^n, we can compute the value of D_n precisely. For any $m \le n$,

$$\begin{aligned}
&\mathbb{E}[J|D_n > m]\,\mathbb{P}(D_n > m) \\
&\quad = \sum_{b_1^n : D_n > m} \mathbb{E}\left[J|B_1^n = b_1^n\right]\mathbb{P}\left(B_1^n = b_1^n\right) \\
&\quad \overset{(a)}{=} \sum_{b_1^n : D_n \ge m+1} \mathbb{E}\left[J|B_1^n = b_1^n\right]\mathbb{P}\left(B_1^n = b_1^n\right) \\
&\quad \overset{(b)}{\le} 4\log\left(\frac{1+\tau_\star}{1-\tau_\star}\right)\cdot \tau_\star^m,
\end{aligned} \tag{103}$$

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

$$\begin{aligned}
&\mathbb{E}[J|D_n \le m]\,\mathbb{P}(D_n \le m) \\
&\quad = \sum_{b_1^n : D_n \le m} \mathbb{E}\left[J|B_1^n = b_1^n\right]\mathbb{P}\left(B_1^n = b_1^n\right) \\
&\quad \overset{(a)}{\le} \log|\mathcal{A}|\cdot\mathbb{P}(D_n \le m) \\
&\quad \overset{(b)}{\le} \log|\mathcal{A}|\cdot\gamma\frac{(n\gamma)^m}{m!}\rho^{n+1},
\end{aligned} \tag{104}$$

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

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

$$\begin{aligned}
&I(A_0; A_{n+1}|B_1^n) = \mathbb{E}[J] \\
&= \mathbb{E}[J|D_n > m]\,\mathbb{P}(D_n > m) + \mathbb{E}[J|D_n \le m]\,\mathbb{P}(D_n \le m) \\
&\le \log\left(\frac{1+\tau_\star}{1-\tau_\star}\right)\cdot\tau_\star^m + (\gamma\cdot\log|\mathcal{A}|)\cdot\frac{(n\gamma)^m}{m!}\rho^{n+1}.
\end{aligned}$$

Denoting $\alpha = \gamma\cdot\log|\mathcal{A}|$ 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 θ is sufficiently small. To this end, we use the inequality $m! \ge (m/e)^m$, see [42, p. 52]. We set $m = \theta n$, and upper-bound the second summand in the right-hand side of (100) to obtain

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

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

That said, taking $m = \theta n$ might not be the best strategy for minimizing n in the right-hand side of (100). A different startegy 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 $(n_\star, \delta_\star, \tau_\star)$-KHMM for some $n_\star, \delta_\star, \tau_\star$. Let

$$m = \left\lceil \log_{\tau_\star}\left(\frac{\epsilon}{2}\cdot\frac{1}{4\log\left(\frac{1+\tau_\star}{1-\tau_\star}\right)}\right)\right\rceil.$$

[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: $\lim_{\theta\to 0}\theta\ln(1/\theta) = \lim_{\theta\to 0}\ln(1/\theta)/(1/\theta) = \lim_{\theta\to 0}(-1/\theta)/(-1/\theta^2) = \lim_{\theta\to 0}\theta = 0$.

Case 1: If $\delta_\star = 0$ then at time $\lambda = (m+1)n_\star$ the sequence B_1^λ can be divided into $m+1$ continguous sequences of length $n_\star$, each inducing a Birkhoff contraction coefficient less than $\tau_\star$. Therefore, using Corollary 32 we obtain that in this case for any $n \ge \lambda$,

$$I(A_0; A_{n+1}|B_1^n) \le 4\log\left(\frac{1+\tau_\star}{1-\tau_\star}\right)\tau_\star^m \le \frac{\epsilon}{2}.$$

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

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

where $\gamma = 1/\delta_\star$, $\alpha = \gamma\cdot\log|\mathcal{A}|$, and $\rho = \delta_\star^{1/n_\star}$. Such λ exists since m is fixed and $\rho < 1$. For this m and any $n \ge \lambda$, the right-hand side of (100) is upper-bounded by ϵ. ■

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 γ and ρ are conservative.

A given KHMM may be associated with many combinations of $(n_\star, \delta_\star, \tau_\star)$. Thus, one needs to carefully select the right combination of these parameters to minimize λ in Theorem 33. A more refined analysis, that considers a KHMM for which multiple combinations $(n_\star, \delta_\star, \tau_\star)$ 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 λ. 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, Arıkan 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 $(\eta, \mathcal{L}, \mathcal{H})$-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 Arıkan'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*

$$Z_{n+1} \le \kappa \cdot \begin{cases} Z_n^2, & B_{n+1} = 0, \\ Z_n, & B_{n+1} = 1, \end{cases} \qquad n \ge 0, \tag{105}$$

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

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

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[12] 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 Bhattcharyya parameter with $\kappa = 2$ was established in [2, Proposition 5]. Under memory, (105) was shown in [13, Theorem 2], with $\kappa = 2\psi_0$, where ψ_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

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

The magnitude of n_0 depends on two factors: the almost-sure convergence of Z_n to Z_∞ 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 $\epsilon_a > 0$ and $\delta_a > 0$ there must be an n_a such that

$$\mathbb{P}\left(Z_n \le \epsilon_a\right) \ge \mathbb{P}\left(Z_\infty = 0\right) - \delta_a, \quad \forall n \ge n_a. \tag{107}$$

We reiterate that n_a 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[13] — is universally smaller than η, which can be set as small as desired.[14] The ability to set η 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 $Z_0 \le \eta$ then*

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

Crucially, η and n_0 depend on the process Z_n only through κ. Inspection of the proof of [43] reveals that Lemma 40 will be true once it is shown that for any $\epsilon_a > 0$ and $\delta' > 0$ there exists n_a such that

$$\mathbb{P}\left(Z_n \le \epsilon_a \text{ for all } n \ge n_a\right) \ge 1 - \delta'. \tag{109}$$

The crux of our proof will be to show that we can set $\eta > 0$ and n_a 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 Arıkan'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

$$-b \le T \le a,$$

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

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

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

$$\mathcal{A}_\alpha(n) = \{J_m \ge \alpha \text{ for some } m \le n\}$$

and

$$\mathcal{A}_\alpha = \{J_m \ge \alpha \text{ for some } m \in \mathbb{N}\}.$$

[12] In which (106) is replaced with $\lim_{n_0\to\infty} \mathbb{P}(\forall n \ge n_0,\ Z_n \le 2^{-2^{n\beta}}) = \mathbb{P}(Z_\infty = 0)$.

[13] See [14, Lemma 1] for relationships between the Bhattcharyya parameter and the conditional entropy.

[14] More 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 $\mathcal{A}_\alpha(n) \subseteq \mathcal{A}_\alpha(n+1)$ and $\cup_{n=1}^{\infty}\mathcal{A}_\alpha(n) = \mathcal{A}_\alpha$, so that by continuity of measure [37, Theorem 2.1],

$$\mathbb{P}(\mathcal{A}_\alpha) = \lim_{n\to\infty} \mathbb{P}(\mathcal{A}_\alpha(n)). \tag{111}$$

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

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

$$\mathbb{P}(\mathcal{A}_\alpha) \le e^{-r\alpha}. \tag{112}$$

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

$$\mathbb{P}(J_n < n(1-\gamma)\mu) \ge 1 - e^{-2n\left(\frac{\gamma\mu}{a+b}\right)^2}. \tag{113}$$

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 $n_a \in \mathbb{N}$ and $\alpha > 0$ such that with probability arbitrarily close to 1, J_n drops below that threshold for every $n \ge n_a$ and never (for any $n \in \mathbb{N}$) visits above α. 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,

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

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,

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

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

$$g(s) \ge \mathbb{E}\left[e^{sT} \cdot \mathbf{1}\{T \ge a/2\}\right] \ge e^{sa/2}p.$$

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

$$r \triangleq \max_{s>0} \left\{s : \mathbb{E}\left[e^{sT}\right] = 1\right\}. \tag{114}$$

For the r found above, denote

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

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

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

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

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

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

$$\tilde{J}_{n\wedge\tau} \triangleq \begin{cases}\tilde{J}_n, & \tau > n,\\ \tilde{J}_\tau, & \tau \le n\end{cases}$$

is also a martingale, and

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

Observe that for any $n \in \mathbb{N}$, we have $\mathbb{P}(\mathcal{A}_\alpha(n)) = \mathbb{P}(\tau \le n)$. Thus,

$$\begin{aligned}1 &= \mathbb{E}\left[\tilde{J}_{n\wedge\tau}\right]\\ &= \mathbb{E}\left[\tilde{J}_{n\wedge\tau} | \tau \le n\right] \cdot \mathbb{P}(\mathcal{A}_\alpha(n))\\ &\quad + \mathbb{E}\left[\tilde{J}_{n\wedge\tau} | \tau > n\right] \cdot \Big(1 - \mathbb{P}(\mathcal{A}_\alpha(n))\Big)\\ &\overset{\text{(a)}}{\ge} \mathbb{E}\left[\tilde{J}_{n\wedge\tau} | \tau \le n\right] \mathbb{P}(\mathcal{A}_\alpha(n))\\ &\overset{\text{(b)}}{=} \mathbb{E}\left[\tilde{J}_\tau | J_\tau \ge \alpha, \tau \le n\right] \mathbb{P}(\mathcal{A}_\alpha(n))\\ &= \mathbb{E}\left[e^{rJ_\tau} | J_\tau \ge \alpha, \tau \le n\right] \mathbb{P}(\mathcal{A}_\alpha(n))\\ &\overset{\text{(c)}}{\ge} e^{r\alpha}\mathbb{P}(\mathcal{A}_\alpha(n)).\end{aligned}$$

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

$$\mathbb{P}(\mathcal{A}_\alpha(n)) \le e^{-r\alpha}.$$

Thus, by (111),

$$\mathbb{P}(\mathcal{A}_\alpha) = \lim_{n\to\infty} \mathbb{P}(\mathcal{A}_\alpha(n)) \le e^{-r\alpha}.$$

This completes the proof of (112).

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

$$\mathbb{P}(J_n \ge n(\mu + t)) \le e^{-2n\left(\frac{t}{a+b}\right)^2}.$$

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

$$\begin{aligned}\mathbb{P}(J_n < n(1-\gamma)\mu) &= 1 - \mathbb{P}(J_n \ge n(\mu + \gamma|\mu|))\\ &\ge 1 - e^{-2n\left(\frac{\gamma\mu}{a+b}\right)^2}.\end{aligned}$$

This completes the proof. ∎

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

$$\begin{aligned}&\mathbb{P}\left(\{\forall n \ge n_a,\ J_n < n_a(1-\gamma)\mu\} \cap \mathcal{A}_\alpha^{\mathrm{c}}\right)\\ &\quad \ge 1 - \left(1 - e^{-2\left(\frac{\gamma\mu}{a+b}\right)^2}\right)^{-1} \cdot e^{-2n_a\left(\frac{\gamma\mu}{a+b}\right)^2} - e^{-r\alpha}.\end{aligned} \tag{115}$$

Proof: Note that

$$
\begin{aligned}
&\mathbb{P}(\forall n \ge n_a,\ J_n < n_a(1-\gamma)\mu) \\
&\quad = \mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n_a(1-\gamma)\mu\}\right) \\
&\quad \ge \mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n(1-\gamma)\mu\}\right) \\
&\quad = 1 - \mathbb{P}\left(\bigcup_{n=n_a}^{\infty} \{J_n \ge n(1-\gamma)\mu\}\right) \\
&\quad \overset{(a)}{\ge} 1 - \sum_{n=n_a}^{\infty} e^{-2n\left(\frac{\gamma\mu}{a+b}\right)^2} \\
&\quad = 1 - \left(\frac{1}{1-e^{-2\left(\frac{\gamma\mu}{a+b}\right)^2}}\right) \cdot e^{-2n_a\left(\frac{\gamma\mu}{a+b}\right)^2}, \qquad (116)
\end{aligned}
$$

where (a) is by (113) and the union bound. Observing that

$$
\begin{aligned}
&\mathbb{P}(\forall n \ge n_a,\ J_n < n(1-\gamma)\mu) \\
&\quad = \mathbb{P}(\{\forall n \ge n_a,\ J_n < n(1-\gamma)\mu\} \cap \mathcal{A}_\alpha) \\
&\qquad + \mathbb{P}\left(\{\forall n \ge n_a,\ J_n < n(1-\gamma)\mu\} \cap \mathcal{A}_\alpha^{\mathrm{c}}\right) \\
&\quad \le \mathbb{P}(\mathcal{A}_\alpha) + \mathbb{P}\left(\{\forall n \ge n_a,\ J_n < n(1-\gamma)\mu\} \cap \mathcal{A}_\alpha^{\mathrm{c}}\right),
\end{aligned}
$$

we obtain

$$
\begin{aligned}
&\mathbb{P}\left(\{\forall n \ge n_a,\ J_n < n(1-\gamma)\mu\} \cap \mathcal{A}_\alpha^{\mathrm{c}}\right) \\
&\qquad \ge \mathbb{P}(\forall n \ge n_a,\ J_n < n(1-\gamma)\mu) - \mathbb{P}(\mathcal{A}_\alpha).
\end{aligned}
$$

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 η such that if $Z_0 \le \eta$, then (109) holds. Thus, we fix $\epsilon_a > 0$ and $\delta' > 0$, and work toward this goal.

Let the process $\bar{Z}_0, \bar{Z}_1, \ldots$ be defined as

$$
\begin{aligned}
\bar{Z}_0 &= \ln Z_0, \\
\bar{Z}_{n+1} &= \begin{cases} 2\bar{Z}_n + \ln\kappa, & B_{n+1} = 0, \\ \bar{Z}_n + \ln\kappa, & B_{n+1} = 1, \end{cases} \quad n \ge 0. \qquad (117)
\end{aligned}
$$

Then, by (105), $\ln Z_n \le \bar{Z}_n$ for any n. Therefore, (109) will be true once we show that there exists n_a and η such that if $\bar{Z}_0 = \ln\eta$, then

$$\mathbb{P}\left(\bar{Z}_n \le \ln\epsilon_a \text{ for all } n \ge n_a\right) \ge 1 - \delta'.$$

Fix

$$0 < \zeta < 1/\kappa^2 \qquad (118)$$

such that $\bar{Z}_0 < \ln\zeta < 0$. Since $\bar{Z}_0 = \ln\eta$ by assumption, and since we may set η as small as desired, we can ensure that this is possible. We then have, by (105),

$$\bar{Z}_1 \le \begin{cases} \bar{Z}_0 + \ln\kappa + \ln\zeta, & B_n = 0, \\ \bar{Z}_0 + \ln\kappa, & B_n = 1. \end{cases}$$

If, further, $\bar{Z}_1 < \ln\zeta$ then the above inequality holds when $\bar{Z}_1$ and $\bar{Z}_0$ are replaced with $\bar{Z}_2$ and $\bar{Z}_1$, respectively. More generally, we define the process J_n, $n \in \mathbb{N}$, by

$$
\begin{aligned}
J_0 &= \bar{Z}_0 = \ln\eta, \\
J_{n+1} &= J_n + T_{n+1}, \quad n \ge 0,
\end{aligned}
$$

where

$$T_n = \begin{cases} \ln\kappa + \ln\zeta, & B_n = 0, \\ \ln\kappa, & B_n = 1, \end{cases} \quad n \ge 1.$$

If $J_i < \ln\zeta$ for all $i \le n$, then $\bar{Z}_n \le J_n$.

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:

$$-\ln\left(\frac{1}{\kappa\zeta}\right) = -b \le T \le a = \ln\kappa.$$

Both a and b are positive by (118) and since $\kappa > 1$ by assumption. By definition, for any $\epsilon > 0$, $\mathbb{P}(T > a - \epsilon) \ge \mathbb{P}(T = a) = 1/2$. 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

$$\mathbb{E}\left[e^{rT}\right] = \frac{(\kappa\zeta)^r + \kappa^r}{2} = 1. \qquad (119)$$

Such r exists, as shown in the proof of Lemma 41. Denote for brevity

$$\theta \triangleq \left|\frac{\mu}{a+b}\right|.$$

By Corollary 42, for any $0 < \gamma < 1$ and $n_a \ge 0$ we have

$$
\begin{aligned}
&\mathbb{P}\left(\{\forall n \ge n_a,\ J_n - J_0 < n_a(1-\gamma)\mu\} \cap \mathcal{A}^{\mathrm{c}}_{-J_0+\ln\zeta}\right) \\
&\overset{(a)}{=} \mathbb{P}\left(\{\forall n \ge n_a,\ J_n < J_0 - n_a(1-\gamma)|\mu|\} \cap \mathcal{A}^{\mathrm{c}}_{-J_0+\ln\zeta}\right) \\
&\ge 1 - (1-e^{-2\gamma^2\theta^2})^{-1} e^{-2n_a\gamma^2\theta^2} - e^{-r(-J_0+\ln\zeta)}, \qquad (120)
\end{aligned}
$$

where (a) is because $\mu < 0$.

Observe that since $J_n = J_0 + \sum_{i=1}^{n} T_i$ we have

$$
\begin{aligned}
\mathcal{A}^{\mathrm{c}}_{-J_0+\ln\zeta} &= \left\{\sum_{i=1}^{n} T_i < -J_0 + \ln\zeta \text{ for all } n \in \mathbb{N}\right\} \\
&= \{J_n < \ln\zeta \text{ for all } n \in \mathbb{N}\}.
\end{aligned}
$$

Consequently, under the event $\mathcal{A}^{\mathrm{c}}_{-J_0+\ln\zeta}$, we have $\bar{Z}_n \le J_n$ for any n. Hence,

$$\mathbb{P}\left(\{\forall n \ge n_a,\ J_n < J_0 - n_a(1-\gamma)|\mu|\} \cap \mathcal{A}^{\mathrm{c}}_{-J_0+\ln\zeta}\right)$$

lower-bounds the probability that $\bar{Z}_n \le J_0 + n_a(1-\gamma)\mu$ for all $n \ge n_a$.

Recall that $J_0 = \bar{Z}_0 = \ln\eta$. It remains to set η and n_a such that $\ln\eta < \ln\zeta$, $J_0 - n_a(1-\gamma)|\mu| \le \ln\epsilon_a$, and the right-hand side of (120) exceeds $1 - \delta'$. Below we show one selection of

η and n_a. Observe that there is freedom in this selection, and generally it is desirable to find small n_a and large η. We leave such optimization for future work.

We first set the parameters γ and ζ. We take $\gamma = 1/2$ and $\zeta = 1/(2\kappa^2)$. In this case, $|\mu| = (\ln 2)/2$ and $\theta = \ln 2/(2\ln(2\kappa^2))$. Further, our plan is to split δ' 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 ζ into (119) and compute r, the largest positive solution of $\kappa^r + (2\kappa)^{-r} = 2$.

Next, we set J_0 so that $e^{-r(-J_0+\ln\zeta)} \le \delta'/2$; one choice is $J_0 = \ln\zeta + \frac{1}{r}\ln(\delta'/2)$. Observe that indeed $J_0 = \ln\eta < \ln\zeta$ since $\delta' < 1$ (there is nothing to prove if $\delta' \ge 1$). We thus take

$$\eta = e^{J_0} = \frac{1}{2\kappa^2}\left(\frac{\delta'}{2}\right)^{1/r}.$$

We set n_a large enough such that both $J_0 - n_a|\mu|/2 \le \ln\epsilon_a$ and $(1-e^{-2\gamma^2\theta^2})^{-1}e^{-2n_a\gamma^2\theta^2} \le \delta'/2$. That is, $n_a = \lceil n'_a \rceil$, where

$$n'_a = \max\left\{\frac{4}{\ln 2}(J_0 - \ln\epsilon_a), \frac{2}{\theta^2}\ln\left(\frac{2}{\delta'\cdot(1-e^{-\theta^2/2})}\right)\right\}.$$

For the above η and n_a, $\mathbb{P}(\bar{Z}_n \le \ln\epsilon_a \text{ for all } n \ge n_a) \ge 1-\delta'$. Thus, (109) holds, and the proof is complete. ∎

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

The values of n_a and η 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 η and numerically run the recursion (117) until $\bar{Z}_n$ is sufficiently small for most indices. The above proof implies that if η 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 $T_M^N = (X_M^N, Y_M^N)$ with realization t_M^N. For brevity, we denote $P_{T_M^N} = P_{T_M^N}(t_M^N)$, and similarly $P_{S_N} = P_{S_N}(s_N)$.

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

$$\psi_k = \begin{cases} \max\limits_{s,\sigma} \dfrac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)\,\mathbb{P}(S_k = \sigma)}, & k > 0, \\ \max\limits_{s} \dfrac{1}{\mathbb{P}(S_0 = s)}, & k = 0 \end{cases}$$

and

$$\phi_k = \begin{cases} \min\limits_{s,\sigma} \dfrac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)\,\mathbb{P}(S_k = \sigma)}, & k > 0, \\ 0, & k = 0. \end{cases}$$

Recall that by stationarity, $P_{S_0} = P_{S_k}$ for any k. Further, observe that by Bayes' law,

$$\frac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)\,\mathbb{P}(S_k = \sigma)} = \frac{\mathbb{P}(S_k = \sigma | S_0 = s)}{\mathbb{P}(S_k = \sigma)}.$$

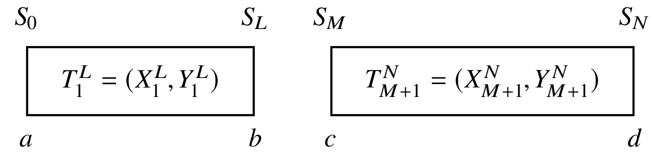

Fig. 11. Two blocks of a FAIM process, not necessarily of the same length. The state S_0, just before the first block, assumes value $a \in \mathcal{S}$. The final state of the first block, S_L, assumes value $b \in \mathcal{S}$. The state S_M, just before the second block, assumes value $c \in \mathcal{S}$. The final state of the second block, S_N, assumes value $d \in \mathcal{S}$.

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 Figure 11). Then,

$$\begin{aligned}
P_{T_1^L, T_{M+1}^N} &= \sum_{t_{L+1}^M} P_{T_1^L, T_{L+1}^M, T_{M+1}^N} \\
&= \sum_{t_{L+1}^M} \sum_{d,a} P_{T_1^L, T_{L+1}^M, T_{M+1}^N, S_N | S_0} P_{S_0} \\
&= \sum_{\substack{d,c,\\ b,a}} \sum_{t_{L+1}^M} P_{T_{M+1}^N, S_N | S_M} P_{T_{L+1}^M, S_M | S_L} P_{T_1^L, S_L | S_0} P_{S_0} \\
&= \sum_{\substack{d,c,\\ b,a}} P_{T_{M+1}^N, S_N | S_M} \left(\sum_{t_{L+1}^M} P_{T_{L+1}^M, S_M | S_L}\right) P_{T_1^L, S_L | S_0} P_{S_0} \\
&= \sum_{\substack{d,c,\\ b,a}} P_{T_{M+1}^N, S_N | S_M} P_{S_M | S_L} P_{T_1^L, S_L | S_0} P_{S_0} \\
&= \sum_{\substack{d,c,\\ b,a}} P_{T_{M+1}^N, S_N | S_M} P_{S_M} \frac{P_{S_M | S_L}}{P_{S_M}} P_{T_1^L, S_L | S_0} P_{S_0} \\
&\overset{(a)}{\le} \psi_{M-L} \left(\sum_{d,c} P_{T_{M+1}^N, S_N | S_M} P_{S_M}\right) \left(\sum_{b,a} P_{T_1^L, S_L | S_0} P_{S_0}\right) \\
&= \psi_{M-L} P_{T_1^L} P_{T_{M+1}^N},
\end{aligned}$$

where (a) follows from the definition of ψ_k. This shows (33a). To see (33b) we follow the exact steps above up to just before inequality (a), and proceed with

$$\begin{aligned}
P_{T_1^L, T_{M+1}^N} &\ge \phi_{M-L} \left(\sum_{d,c} P_{T_{M+1}^N, S_N | S_M} P_{S_M}\right) \left(\sum_{b,a} P_{T_1^L, S_L | S_0} P_{S_0}\right) \\
&= \phi_{M-L} P_{T_1^L} P_{T_{M+1}^N}.
\end{aligned}$$

Again, the inequality follows from the definition of ϕ_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,

$$\begin{aligned}
P_{T_1^L, T_{L+1}^N} &= \sum_{\substack{a,b,\\ d}} P_{T_{L+1}^N, S_N | S_L} \frac{P_{S_L}}{P_{S_L}} P_{T_1^L, S_L | S_0} P_{S_0} \\
&\le \psi_0 \sum_{d,b} P_{T_{L+1}^N, S_N | S_L} P_{S_L} \cdot \left(\sum_{b',a} P_{T_1^L, S_L | S_0} P_{S_0}\right) \\
&= \psi_0 P_{T_1^L} P_{T_{L+1}^N};
\end{aligned}$$

where the inequality is by the definition of ψ_0 and because $P_{T_1^L, S_L|S_0} \leq \sum_{b'} P_{T_1^L, S_L|S_0}$.

To see that that ψ_k is nonincreasing, observe that for any $s, \sigma \in \mathcal{S}$:

$$\begin{aligned} P_{S_{k+1}, S_0}(\sigma, s) &= \sum_{a \in \mathcal{S}} P_{S_{k+1}|S_k}(\sigma|a) \cdot P_{S_k, S_0}(a, s) \\ &\leq \psi_k \sum_{a \in \mathcal{S}} P_{S_{k+1}|S_k}(\sigma|a) \cdot P_{S_k}(a) P_{S_0}(s) \\ &= \psi_k P_{S_{k+1}}(\sigma) P_{S_0}(s). \end{aligned}$$

Therefore, we must have $\psi_{k+1} \leq \psi_k$. The proof that ϕ_k is nondecreasing is similar, with "$\leq \psi_k$" replaced with "$\geq \phi_k$".

Finally, the asymptotic properties of ϕ_k and ψ_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

$$\left| \frac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)\,\mathbb{P}(S_k = \sigma)} - 1 \right| \leq \psi_0 \cdot \alpha \cdot \gamma^k \xrightarrow[k \to \infty]{} 0.$$

Hence, both ψ_k and ϕ_k must tend to 1 exponentially fast as $k \to \infty$. ■

Proof of Lemma 7: The FAIM process is forgetful, so we let λ be the ϵ-recollection of the process. For this λ, (34) is satisfied.

By the chain rule for mutual information,

$$\begin{aligned} &I(S_0; S_{-k}, S_k | X_{-\ell}^{-1}, Y_{-\ell}^{m}) \\ &\quad = I(S_0; S_k | X_{-\ell}^{-1}, Y_{-\ell}^{m}) + I(S_0; S_{-k} | X_{-\ell}^{-1}, Y_{-\ell}^{m}, S_k). \end{aligned} \tag{121}$$

We will upper-bound each of the terms on the right-hand side of (121) by ϵ, yielding the desired result.

For any m, ℓ, k such that $\min\{m, \ell\} \geq k \geq \lambda$ we have

$$\begin{aligned} \epsilon &\overset{(a)}{\geq} I(S_0; S_k | Y_0^k) \\ &\overset{(b)}{\geq} I(S_0; (S_k, Y_{k+1}^m) | Y_0^k) \\ &= I(S_0; Y_{k+1}^m | Y_0^k) + I(S_0; S_k | Y_0^m) \\ &\overset{(c)}{\geq} I(S_0; S_k | Y_0^m) \\ &\overset{(d)}{\geq} I((S_0, X_{-\ell}^{-1}, Y_{-\ell}^{-1}); S_k | Y_0^m) \\ &= I(X_{-\ell}^{-1}, Y_{-\ell}^{-1}; S_k | Y_0^m) + I(S_0; S_k | X_{-\ell}^{-1}, Y_{-\ell}^m) \\ &\overset{(e)}{\geq} I(S_0; S_k | X_{-\ell}^{-1}, Y_{-\ell}^m). \end{aligned}$$

We now justify the inequalities:[15]

- (a) is by (34b) and stationarity.
- (b) is by (2), noting that (32) implies
$$S_0 -\!\!\circ\!\!- (S_k, Y_0^k) -\!\!\circ\!\!- (S_k, Y_{k+1}^m);$$
- (c) is because mutual information is nonnegative;
- (d) is by (2), noting that (32) implies
$$S_k -\!\!\circ\!\!- (S_0, Y_0^m) -\!\!\circ\!\!- (S_0, X_{-\ell}^{-1}, Y_{-\ell}^{-1})$$
(observe that $X_{-\ell}^{-1}, Y_{-\ell}^{-1}$ 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, ℓ, k such that $\min\{m, \ell\} \geq k \geq \lambda$ we have

$$\begin{aligned} \epsilon &\overset{(a)}{\geq} I(S_0; S_{-k} | X_{-k}^{-1}, Y_{-k}^{-1}) \\ &\overset{(b)}{\geq} I(S_0; (S_{-k}, X_{-\ell}^{-k-1}, Y_{-\ell}^{-k-1}) | X_{-k}^{-1}, Y_{-k}^{-1}) \\ &\overset{(c)}{\geq} I(S_0; S_{-k} | X_{-\ell}^{-1}, Y_{-\ell}^{-1}) \\ &\overset{(d)}{\geq} I((S_0, Y_0^m, S_k); S_{-k} | X_{-\ell}^{-1}, Y_{-\ell}^{-1}) \\ &\overset{(e)}{\geq} I(S_0; S_{-k} | X_{-\ell}^{-1}, Y_{-\ell}^m, S_k). \end{aligned}$$

Again, we justify the inequalities:

- (a) is by (35a) and stationarity.
- (b) is by (2), noting that (32) implies
$$S_0 -\!\!\circ\!\!- (S_{-k}, X_{-k}^{-1}, Y_{-k}^{-1}) -\!\!\circ\!\!- (S_{-k}, X_{-\ell}^{-k-1}, Y_{-\ell}^{-k-1});$$
- (c) is by the chain rule for mutual information
- (d) is by (2), noting that (32) implies
$$S_{-k} -\!\!\circ\!\!- (S_0, X_{-\ell}^{-1}, Y_{-\ell}^{-1}) -\!\!\circ\!\!- (S_0, Y_0^m, S_k);$$
- (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 λ as the ϵ-recollection of the process. The corollary holds for $k = 1$ by Lemma 7. We proceed by induction. Assume that the collolary holds for $k - 1 \geq 1$, and we will show it holds for k.

Let

$$\begin{aligned} \mathbf{i}' &= \begin{bmatrix} i_1 & i_2 & \cdots & i_{k-1} \end{bmatrix}, \\ \mathbf{i} &= \begin{bmatrix} i_1 & i_2 & \cdots & i_{k-1} & i_k \end{bmatrix} = \begin{bmatrix} \mathbf{i}' & i_k \end{bmatrix}. \end{aligned}$$

For brevity, denote

$$C_i = (X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}).$$

Our goal is thus to show that

$$\begin{aligned} &I(S_{\mathbf{i}}; S_{\mathbf{i}-L_0}, S_{\mathbf{i}+L_0} | C_{\mathbf{i}}) \\ &\quad = I(S_{\mathbf{i}'}, S_{i_k}; S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, S_{i_k-L_0}, S_{i_k+L_0} | C_{\mathbf{i}'}, C_{i_k}) \leq k \cdot 2\epsilon. \end{aligned}$$

Indeed,

$$\begin{aligned} &I(S_{\mathbf{i}'}, S_{i_k}; S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, S_{i_k-L_0}, S_{i_k+L_0} | C_{\mathbf{i}'}, C_{i_k}) \\ &\quad = I(S_{\mathbf{i}'}; S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, S_{i_k-L_0}, S_{i_k+L_0} | C_{\mathbf{i}'}, C_{i_k}) \\ &\qquad + I(S_{i_k}; S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, S_{i_k-L_0}, S_{i_k+L_0} | S_{\mathbf{i}'}, C_{\mathbf{i}'}, C_{i_k}) \\ &\quad \overset{(a)}{\leq} I(S_{\mathbf{i}'}; S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, (S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k}) | C_{\mathbf{i}'}) \\ &\qquad + I(S_{i_k}; S_{i_k-L_0}, S_{i_k+L_0}, (S_{\mathbf{i}'}, S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, C_{\mathbf{i}'}) | C_{i_k}) \\ &\quad \overset{(b)}{\leq} I(S_{\mathbf{i}'}; S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0} | C_{\mathbf{i}'}) + I(S_{i_k}; S_{i_k-L_0}, S_{i_k+L_0} | C_{i_k}) \\ &\quad \overset{(c)}{\leq} (k-1) \cdot 2\epsilon + 2\epsilon \\ &\quad = k \cdot 2\epsilon, \end{aligned}$$

[15] We remark that in (b) and (d) we can replace the inequalities with equalities.

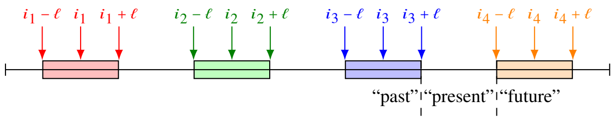

Fig. 12. Illustration of the timeline for $k = 4$. Given $S_{i_4-\ell}$, the “future” is independent of the “present” and “past.” Given $S_{i_3+\ell}$, the “past” is independent of the “present” and “future.”

where (a) is by the chain rule; (b) is by (2) and (32), used for the Markov chains (See Figure 12 for an illustration):

$$S_{\mathbf{i}'} -\!\!\circ\!\!- (S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, C_{\mathbf{i}'}) -\!\!\circ\!\!- (S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k}),$$

which holds because $i_{k-1} \le i_{k-1} + L_0 \le i_k - L_0$ so $(S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k})$ are independent of $S_{\mathbf{i}'}$ given $S_{i_{k-1}+L_0}$, which is part of $S_{\mathbf{i}'+L_0}$, and

$$\begin{aligned} S_{i_k} &-\!\!\circ\!\!- (S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k}) \\ &-\!\!\circ\!\!- (S_{i_k-L_0}, S_{i_k+L_0}, S_{\mathbf{i}'}, S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, C_{\mathbf{i}'}), \end{aligned}$$

which again holds because $i_{k-1} + L_0 \le i_k - L_0 \le i_k$, so $(S_{\mathbf{i}'}, S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, C_{\mathbf{i}'})$ are indepndent of S_{i_k} given $S_{i_{k-L_0}}$; finally, (c) is because $I(S_{\mathbf{i}'}; S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0} | C_{\mathbf{i}'}) \le (k-1)\cdot 2\epsilon$ by the induction hypothesis and $I(S_{i_k}; S_{i_k-L_0}, S_{i_k+L_0} | C_{i_k}) \le 2\epsilon$ by Lemma 7. This completes the proof. ■

Appendix C
Auxiliary Proofs for Section IV-B

Recall from (1) that the binary entropy function $h_2 : [0,1] \to [0,1]$ is defined by

$$h_2(x) = -x \log x - (1-x)\log(1-x).$$

This is a concave-∩ function that satisfies $h_2(x) = h_2(1-x)$ for any $x \in [0,1]$, and is monotone increasing over $[0,1/2]$. The inverse of the binary entropy function is $h_2^{-1} : [0,1] \to [0,1/2]$. The following three technical lemmas will be used to prove Lemma 13.

Lemma 43. *For any $0 \le x \le 1/2$,*

$$1 - h_2(x) \ge \frac{2}{\ln 2}\left(\frac{1}{2} - x\right)^2. \tag{122}$$

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, $1 - h_2(x) \ge 2(1/2 - x)^2/(\ln 2)$ for any $0 \le x \le 1/2$ as well. ■

Lemma 44. *For any $0 \le y \le x \le 1/2$,*

$$h_2(x) - h_2(y) \ge \frac{1}{\ln 2}(x-y)(1-2y). \tag{123}$$

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 $x_1 \le x_2 \le x_3$ we have

$$(x_3 - x_1)(h_2(x_2) - h_2(x_1)) \ge (x_2 - x_1)(h_2(x_3) - h_2(x_1)) \tag{124}$$

(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

$$\left(\frac{1}{2} - y\right)(h_2(x) - h_2(y)) \ge (x-y)(1 - h_2(y)).$$

Since $y < x \le 1/2$ by assumption, $1/2 - y > 0$. Therefore, we rearrange the above inequality and obtain

$$h_2(x) - h_2(y) \ge (x-y)\frac{1 - h_2(y)}{1/2 - y} \ge \frac{1}{\ln 2}(x-y)(1-2y),$$

where the rightmost inequality is by (122). ■

Lemma 45. *For any $x, y \in (0, 1/2)$, the function*

$$f(x,y) = h_2(h_2^{-1}(x) * h_2^{-1}(y)) - y \tag{125}$$

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 fuction

$$\operatorname{arctanh}(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right),$$

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, $\operatorname{arctanh}(x) > 0$ for $x > 0$.

Now,

$$\frac{\partial g(x,y)}{\partial x} = \frac{2}{\ln 2}(1-2y)\operatorname{arctanh}\big((1-2x)(1-2y)\big).$$

This is positive since $\operatorname{arctanh}(z) > 0$ 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,

$$\begin{aligned} &\frac{\partial g(x,y)}{\partial y} \\ &= \frac{2}{\ln 2}\Big((1-2x)\operatorname{arctanh}\big((1-2x)(1-2y)\big) - \operatorname{arctanh}(1-2y)\Big) \\ &\le \frac{2}{\ln 2}\Big((1-2x)\operatorname{arctanh}(1-2y) - \operatorname{arctanh}(1-2y)\Big) \\ &= \frac{2}{\ln 2}\left((1-2x) - 1\right)\cdot \operatorname{arctanh}(1-2y) \\ &< 0, \end{aligned}$$

where the first inequality is because $(1-2x)(1-2y) < (1-2y)$ and $\operatorname{arctanh}(\cdot)$ 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

$$\sum_{a,b} p_a q_b h_2(\alpha_a * \beta_b) \ge h_2(h_2^{-1}(A) * h_2^{-1}(B)),$$

where
$$A=\sum_a p_a h_2(\alpha_a), \quad B=\sum_b q_b h_2(\beta_b).$$
Therefore,
$$\sum_{a,b} p_a q_b \left(h_2(\alpha_a*\beta_b)-h_2(\beta_b)\right) \geq h_2(h_2^{-1}(A)*h_2^{-1}(B))-B$$
$$= f(A,B),$$
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
$$\sum_{a,b} p_a q_b \left(h_2(\alpha_a*\beta_b)-h_2(\beta_b)\right) \geq h_2(h_2^{-1}(\xi_1)*h_2^{-1}(\xi_2))-\xi_2.$$
Define, therefore,
$$\Delta(\xi_1,\xi_2) \triangleq h_2(h_2^{-1}(\xi_1)*h_2^{-1}(\xi_2))-\xi_2. \tag{126}$$
It remains to show that $\Delta(\xi_1,\xi_2)>0$.

To this end, observe that for any $x,y\in(0,1/2)$,
$$h_2(x*y)-h_2(y) \overset{(a)}{\geq} \frac{1}{\ln 2}(x*y-y)\cdot(1-2y)$$
$$= \frac{1}{\ln 2}x(1-2y)^2.$$
where (a) is by (123). Therefore,
$$\Delta(\xi_1,\xi_2) \geq \frac{1}{\ln 2} h_2^{-1}(\xi_1)\left(1-2h_2^{-1}(\xi_2)\right)^2 > 0. \qquad \blacksquare$$
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
$$p(f_0,g_0) \triangleq \mathbb{P}\left(F=f_0, G=g_0\right)=\sum_{\substack{a:f(a)=f_0,\\ g(a)=g_0}} p(a), \tag{127a}$$
$$q(f_0,g_0) \triangleq \mathbb{P}\left(\tilde{F}=f_0, \tilde{G}=g_0\right)=\sum_{\substack{a:f(a)=f_0,\\ g(a)=g_0}} q(a). \tag{127b}$$
With this notation we also have $p(g_0)=\mathbb{P}(G=g_0)$ 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 $g_0\in\mathcal{G}$,
$$(1-\varepsilon)q(f_0,g_0) \leq p(f_0,g_0) \leq (1+\varepsilon)q(f_0,g_0),$$
$$(1-\varepsilon)q(g_0) \leq \quad p(g_0) \quad \leq (1+\varepsilon)q(g_0). \tag{128}$$
Therefore,
$$\frac{1-\varepsilon}{1+\varepsilon}\cdot q(f_0|g_0) \leq p(f_0|g_0) \leq \frac{1+\varepsilon}{1-\varepsilon}\cdot q(f_0|g_0).$$
When $0\leq\varepsilon\leq\frac{1}{3}$, we have $(1+\varepsilon)/(1-\varepsilon)\leq 1+3\varepsilon$ and $(1-\varepsilon)/(1+\varepsilon)\geq 1-3\varepsilon\geq 0$ by straightforward algebra. Hence, for any $f_0\in\{0,1\}$ and $g_0\in\mathcal{G}$,
$$(1-3\varepsilon)q(f_0|g_0) \leq p(f_0|g_0) \leq (1+3\varepsilon)q(f_0|g_0),$$
by which $|p(f_0|g_0)-q(f_0|g_0)|\leq 3\varepsilon\cdot q(f_0|g_0)$. Thus, for any $g_0\in\mathcal{G}$, since $\varepsilon<\frac{1}{6}$ by assumption,
$$d(g_0) \triangleq \sum_{f_0=0}^{1} |p(f_0|g_0)-q(f_0|g_0)| \leq 3\varepsilon\sum_{f_0=0}^{1} q(f_0|g_0)=3\varepsilon<\frac{1}{2}.$$
Since F and $\tilde{F}$ are binary, we conclude from [20, Theorem 17.3.3] that for any $g_0\in\mathcal{G}$,
$$\left|H(F|G=g_0)-H(\tilde{F}|\tilde{G}=g_0)\right| \leq -d(g_0)\log\frac{d(g_0)}{2}$$
$$\overset{(a)}{\leq} -3\varepsilon\log\frac{3\varepsilon}{2}. \tag{129}$$
Inequality (a) is true because $x\mapsto -x\log\frac{x}{2}$ is increasing for $0\leq x<\frac{2}{e}\approx 0.736$, and $0\leq d(g_0)\leq 3\varepsilon<\frac{1}{2}<\frac{2}{e}$ by assumption.

Let $\sum^+$ denote summation over all $g_0\in\mathcal{G}$ for which $p(g_0)\geq q(g_0)$, and $\sum^-$ denote summation over all $g_0\in\mathcal{G}$ for which $p(g_0)<q(g_0)$. Since $\sum_{g_0}p(g_0)=\sum_{g_0}q(g_0)=1$, we have
$$\sum{}^{+}(p(g_0)-q(g_0)) = -\sum{}^{-}(p(g_0)-q(g_0))$$
$$= \frac{1}{2}\sum_{g_0}|p(g_0)-q(g_0)|$$
$$\leq \frac{\varepsilon}{2}\sum_{g_0}q(g_0)=\frac{\varepsilon}{2},$$
where the inequality is by (128). Hence, for any nonnegative function $h:\mathcal{G}\to\mathbb{R}^+$,
$$\begin{aligned}&\sum_{g_0}(p(g_0)-q(g_0))h(g_0)\\ &= \sum{}^{+}|p(g_0)-q(g_0)|h(g_0)-\sum{}^{-}|p(g_0)-q(g_0)|h(g_0)\\ &\leq \left(\sup_{g_0}h(g_0)-\inf_{g_0}h(g_0)\right)\cdot\frac{1}{2}\sum_{g_0}|p(g_0)-q(g_0)|\\ &\leq \left(\sup_{g_0}h(g_0)-\inf_{g_0}h(g_0)\right)\cdot\frac{\varepsilon}{2}.\end{aligned} \tag{130}$$
Therefore,
$$\begin{aligned}&H(F|G)-H(\tilde{F}|\tilde{G})\\ &= \sum_{g_0}p(g_0)H(F|G=g_0)-\sum_{g_0}q(g_0)H(\tilde{F}|\tilde{G}=g_0)\\ &\overset{(a)}{\leq} \sum_{g_0}p(g_0)\Big(H(\tilde{F}|\tilde{G}=g_0)-3\varepsilon\log\frac{3\varepsilon}{2}\Big)\\ &\quad -\sum_{g_0}q(g_0)H(\tilde{F}|\tilde{G}=g_0)\\ &= -3\varepsilon\log\frac{3\varepsilon}{2}+\sum_{g_0}(p(g_0)-q(g_0))H(\tilde{F}|\tilde{G}=g_0)\\ &\overset{(b)}{\leq} -3\varepsilon\log\frac{3\varepsilon}{2}+\left(\max_{g_0}H(\tilde{F}|\tilde{G}=g_0)-\min_{g_0}H(\tilde{F}|\tilde{G}=g_0)\right)\cdot\frac{\varepsilon}{2}\\ &\overset{(c)}{\leq} \frac{\varepsilon}{2}-3\varepsilon\log\frac{3\varepsilon}{2},\end{aligned}$$
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,
$$\sum_{g_0}(p(g_0)-q(g_0))h(g_0) \geq -\left(\sup_{g_0}h(g_0)-\inf_{g_0}h(g_0)\right)\cdot\frac{\varepsilon}{2},$$

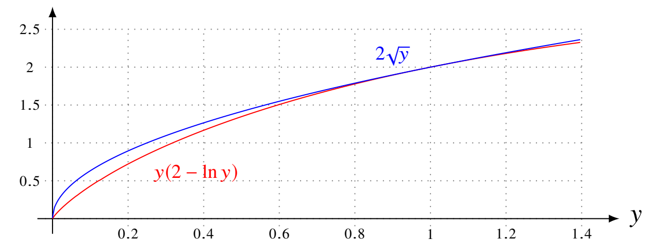

Fig. 13. Illustration of the inequality $y(2-\ln y) \le 2\sqrt{y}$.

by which

$$H(F|G) - H(\tilde{F}|\tilde{G}) \ge -\left(\frac{\varepsilon}{2} - 3\varepsilon \log \frac{3\varepsilon}{2}\right).$$

Thus, we have shown that

$$\left|H(F|G) - H(\tilde{F}|\tilde{G})\right| \le \frac{\varepsilon}{2} - 3\varepsilon \log \frac{3\varepsilon}{2}.$$

By Lemma 46 below and some algebra, we obtain that

$$\frac{\varepsilon}{2} - 3\varepsilon \log \frac{3\varepsilon}{2} \le \sqrt{8} \cdot \frac{2^{1/12}\sqrt{3}}{e \cdot \ln 2}\sqrt{\varepsilon} < \sqrt{8\varepsilon},$$

which completes the proof. ■

Lemma 46. *For any $y > 0$, we have*

$$y(2-\ln y) \le 2\sqrt{y}.$$

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 $xy - e^x \le y \ln y - y$. Now, set $x = \frac{1}{2}\ln y$ and rearrange to yield $y(2-\ln y) \le 2\sqrt{y}$ 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

$$\mathbb{P}\left(B_n = b \,|\, A_n = j, B_1^{n-1}, A_1^{n-1}\right) = \mathbb{P}(B_n = b | A_n = j) = q(b|j). \tag{131}$$

One direction of the equivalence is easy: any deterministic model can be thought of a probabilistic model with $q(\cdot|j)$ 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

$$\begin{aligned}
&\mathbb{P}\left(B_n = b, A_n = j \,|\, A_{n-1} = i, A_1^{n-2}, B_1^{n-1}\right) \\
&\quad = \mathbb{P}(B_n = b, A_n = j \,|\, A_{n-1} = i) \\
&\quad = \mathbb{P}(A_n = j \,|\, A_{n-1} = i) \cdot \mathbb{P}(B_n = b \,|\, A_n = j) \\
&\quad = p(j|i)q(b|j).
\end{aligned}$$

We call a pair (j, b), $j \in \mathcal{A}$, $b \in \mathcal{B}$, *viable* if $q(b|j) > 0$. Define a new Markov chain C_n with states (j, b) whenever (j, b) is a viable pair,[16] and whose transition probability function for any two states (j, b) and (i, k) is $\mathbb{P}(C_n = (j, b) \,|\, C_{n-1} = (i, k)) = p(j|i)q(b|j)$. Set $f : \mathcal{A} \times \mathcal{B} \to \mathcal{B}$ 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 $\mathbb{P}(A_n = j|A_{n-m} = i) > 0$ for all i, j, and n. Therefore, for any viable pairs (j, b) and (i, k),

$$\mathbb{P}(C_n = (j, b)|C_{n-m} = (i, k)) = q(b|j)\mathbb{P}(A_n = j|A_{n-m} = i) > 0.$$

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 $\mathbb{P}(C_n = (j, b)|C_{n-m} = (i, k)) > 0$ for any two viable pairs (states) (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),

$$\mathbb{P}(A_n = j|A_{n-m} = i) = \sum_b \mathbb{P}(C_n = (j, b)|C_{n-m} = (i, k)) > 0.$$

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 proability γ and in the bad state, the channel is a BSC with crossover probability β. 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 detereministic model $C_n = (S_n, X_n, Y_n)$ with states

$$\begin{aligned} 1 &= (\text{good}, 0, 0), & 5 &= (\text{bad}, 0, 0), \\ 2 &= (\text{good}, 0, 1), & 6 &= (\text{bad}, 0, 1), \\ 3 &= (\text{good}, 1, 0), & 7 &= (\text{bad}, 1, 0), \\ 4 &= (\text{good}, 1, 1), & 8 &= (\text{bad}, 1, 1). \end{aligned}$$

For brevity, for a number $x \in [0, 1]$ we denote $\bar{x} = 1 - x$. The transition probability matrix of C_n is

$$\mathsf{M} = \frac{1}{2} \begin{bmatrix} \bar{p}\bar{\gamma} & \bar{p}\gamma & \bar{p}\gamma & \bar{p}\bar{\gamma} & p\bar{\beta} & p\beta & p\beta & p\bar{\beta} \\ \bar{p}\bar{\gamma} & \bar{p}\gamma & \bar{p}\gamma & \bar{p}\bar{\gamma} & p\bar{\beta} & p\beta & p\beta & p\bar{\beta} \\ \bar{p}\bar{\gamma} & \bar{p}\gamma & \bar{p}\gamma & \bar{p}\bar{\gamma} & p\bar{\beta} & p\beta & p\beta & p\bar{\beta} \\ \bar{p}\bar{\gamma} & \bar{p}\gamma & \bar{p}\gamma & \bar{p}\bar{\gamma} & p\bar{\beta} & p\beta & p\beta & p\bar{\beta} \\ q\bar{\gamma} & q\gamma & q\gamma & q\bar{\gamma} & \bar{q}\bar{\beta} & \bar{q}\beta & \bar{q}\beta & \bar{q}\bar{\beta} \\ q\bar{\gamma} & q\gamma & q\gamma & q\bar{\gamma} & \bar{q}\bar{\beta} & \bar{q}\beta & \bar{q}\beta & \bar{q}\bar{\beta} \\ q\bar{\gamma} & q\gamma & q\gamma & q\bar{\gamma} & \bar{q}\bar{\beta} & \bar{q}\beta & \bar{q}\beta & \bar{q}\bar{\beta} \\ q\bar{\gamma} & q\gamma & q\gamma & q\bar{\gamma} & \bar{q}\bar{\beta} & \bar{q}\beta & \bar{q}\beta & \bar{q}\bar{\beta} \end{bmatrix}.$$

The possible observations (X, Y) are $(0, 0), (0, 1), (1, 0)$, and $(1, 1)$. The matrices $\mathsf{M}(b)$, $b \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ are obtained from M by replacing all but two columns of M with zeros. Namely, in $\mathsf{M}(0, 0)$, all but columns 1 and 5 are replaced with zeros; in $\mathsf{M}(0, 1)$ all but columns 2 and 6 are replaced with zeros; in $\mathsf{M}(1, 0)$ all but columns 3 and 7 are replaced with zeros; and in $\mathsf{M}(1, 1)$ all but columns 4 and 8 are replaced with zeros.

Acknowledgment

The authors are grateful to Prof. Rami Atar and Dr. Lele Wang for helpful discussions at the preliminary stages of this work.

References

[1] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, Oct 1998.

[2] E. Arıkan, "Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. on Information Theory*, vol. 55, no. 7, pp. 3051–3073, July 2009.

[3] E. Arıkan and E. Telatar, "On the rate of channel polarization," in *Proc. IEEE Int. Sym. on Information Theory*, June 2009, pp. 1493–1495.

[4] S. H. Hassani, S. B. Korada, and R. Urbanke, "The compound capacity of polar codes," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2009, pp. 16–21.

[5] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacity of a class of channels," *Ann. Math. Statist.*, vol. 30, no. 4, pp. 1229–1241, 12 1959.

[6] E. Şaşoğlu, "Polar Coding Theorems for Discrete Systems," Ph.D. dissertation, School Comput. Commun. Sci., EPFL, Lausanne, Switzerland, 2011. [Online]. Available: https://infoscience.epfl.ch/record/168993/files/EPFL_TH5219.pdf

[7] S. H. Hassani and R. Urbanke, "Universal polar codes," in *2014 IEEE International Symposium on Information Theory*, June 2014, pp. 1451–1455.

[8] E. Şaşoğlu and L. Wang, "Universal polarization," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 2937–2946, June 2016.

[9] D. Sutter and J. M. Renes, "Universal polar codes for more capable and less noisy channels and sources," in *2014 IEEE International Symposium on Information Theory*, June 2014, pp. 1461–1465.

[10] E. Abbe, "Universal source polarization and sparse recovery," in *2010 IEEE Information Theory Workshop*, Aug 2010, pp. 1–5.

[11] M. Ye and A. Barg, "Universal source polarization and an application to a multi-user problem," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2014, pp. 805–812.

[12] M. Alsan, "Universal polar decoding with channel knowledge at the encoder," in *2014 IEEE Information Theory Workshop (ITW 2014)*, Nov 2014, pp. 371–375.

[13] E. Şaşoğlu and I. Tal, "Polar coding for processes with memory," 2016. [Online]. Available: http://arxiv.org/abs/1602.01870

[14] B. Shuval and I. Tal, "Fast polarization for processes with memory," 2017. [Online]. Available: arXiv:1710.02849

[15] R. Wang, R. Liu, and Y. Hou, "Joint successive cancellation decoding of polar codes over intersymbol interference channels," *CoRR*, vol. abs/1404.3001, 2014. [Online]. Available: http://arxiv.org/abs/1404.3001

[16] R. Wang, J. Honda, H. Yamamoto, R. Liu, and Y. Hou, "Construction of polar codes for channels with memory," in *2015 IEEE Information Theory Workshop*, October 2015, pp. 187–191.

[17] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 2, pp. 359–378, May 1994.

[18] B. M. Hochwald and P. R. Jelenković, "State learning and mixing in entropy of hidden Markov processes and the Gilbert-Elliott channel," *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 128–138, Jan 1999.

[19] T. Kaijser, "A limit theorem for partially observed Markov chains," *The Annals of Probability*, vol. 3, no. 4, pp. 677–696, 08 1975.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[21] J. Honda and H. Yamamoto, "Polar coding without alphabet extension for asymmetric models," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 7829–7838, December 2013.

[22] R. Mori and T. Tanaka, "Performance and construction of polar codes on symmetric binary-input memoryless channels," in *Proc. IEEE Int. Sym. on Information Theory*, June 2009, pp. 1496–1500.

[23] I. Tal and A. Vardy, "How to construct polar codes," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6562–6582, October 2013.

[24] R. Pedarsani, S. H. Hassani, I. Tal, and E. Telatar, "On the construction of polar codes," in *2011 IEEE International Symposium on Information Theory Proceedings*, July 2011, pp. 11–15.

[25] I. Tal, A. Sharov, and A. Vardy, "Constructing polar codes for non-binary alphabets and macs," in *2012 IEEE International Symposium on Information Theory Proceedings*, July 2012, pp. 2132–2136.

[26] I. Tal and A. Vardy, "Channel upgrading for semantically-secure encryption on wiretap channels," in *2013 IEEE International Symposium on Information Theory*, July 2013, pp. 1561–1565.

[27] U. Pereg and I. Tal, "Channel upgradation for non-binary input alphabets and macs," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1410–1424, March 2017.

[28] I. Tal, "On the construction of polar codes for channels with moderate input alphabet sizes," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1501–1509, March 2017.

[29] A. Kartowsky and I. Tal, "Greedy-merge degrading has optimal power-law," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1618–1622.

[30] T. C. Gulcu, M. Ye, and A. Barg, "Construction of polar codes for arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 309–321, Jan 2018.

[31] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, 2nd ed. Addison-Wesley, 1994.

[32] M. Iosifescu, *Finite Markov processes and their applications*, ser. Dover books on mathematics. Mineola, N.Y: Dover, 2007.

[33] J. A. Fill, "Eigenvalue bounds on convergence to stationarity for nonreversible markov chains, with an application to the exclusion process," *The Annals of Applied Probability*, vol. 1, no. 1, pp. 62–87, 1991.

[34] E. Seneta, *Non-negative matrices and Markov chains*, ser. Springer series in statistics. New York, NY: Springer, 2006.

[35] J. E. Cohen, "Contractive inhomogeneous products of non-negative matrices," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 86, no. 2, p. 351364, 1979.

[36] M. Artzrouni and X. Li, "A note on the coefficient of ergodicity of a column-allowable nonnegative matrix," *Linear Algebra and its Applications*, vol. 214, pp. 93 – 101, 1995.

[37] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1995.

[38] E. Seneta, "Markov and the birth of chain dependence theory," *International Statistical Review*, vol. 64, no. 3, pp. 255–263, 1996.

[39] F. Kochman and J. Reeds, "A simple proof of Kaijser's unique ergodicity result for hidden Markov α-chains," *The Annals of Applied Probability*, vol. 16, no. 4, pp. 1805–1815, 11 2006.

[40] P. Chigansky and R. van Handel, "A complete solution to Blackwell's unique ergodicity problem for hidden Markov chains," *The Annals of Applied Probability*, vol. 20, no. 6, pp. 2318–2345, 2010.

[41] S. Kullback, "An extension of an information-theoretic derivation of certain limit relations for a Markov chain," *SIAM Journal on Control*, vol. 5, no. 1, pp. 51–53, 1967.

[42] W. Feller, *An Introduction to Probability Theory and Its Applications volume 1*, 3rd ed. Wiley, 1968.

[43] I. Tal, "A simple proof of fast polarization," *IEEE Transactions on Information Theory*, vol. 63, no. 12, pp. 7617–7619, Dec 2017.

[44] H. U. Gerber, "Martingales in risk theory," *Bulletin / Association of Swiss Actuaries*, vol. 73, pp. 205–216, 1973.

[45] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[46] J. S. Rosenthal, *A first look at Rigorous Probability Theory*, 2nd ed. World Scientific, 2006.

[47] D. Williams, *Probability with Martingales*. Cambridge University Press, 1991.

[48] T. M. Apostol, *Mathematical analysis*, 2nd ed. Addison-Wesley, 1974.

[49] D. S. Mitrinović and P. M. Vasić, *Analytic inequalities*, ser. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1970.

[50] J. M. Steele, *The Cauchy-Schwarz Master Class*. Cambridge University Press, 2004.

[51] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[52] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliott channels," *IEEE Transactions on Information Theory*, vol. 35, no. 6, pp. 1277–1290, November 1989.