Universal Polarization for Processes with Memory

Boaz Shuval, 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]. That is, roughly $2^{-\sqrt{\Lambda}}$, where Λ is the codeword length,

The informal statements of the previous paragraph are stated formally in our main theorem, below. The theorem contains several terms that will be defined throughout the paper.

Theorem 1. *Consider a family of FAIM s/o-processes with an upper bound on forgetfulness and mixing, where all s/o-processes share the same input distribution. We consider the case in which the encoder does not know which $s/o-process$ was used, but the decoder does know. Let I^* be the infimal information rate over the family of s/o-processes. Then, for any $R < I^*$ and $\beta < 1/2$, there exists a sequence of codes with growing lengths such that the following holds for each code:*

1) *The code rate is at least R.*
2) *The probability of error is upper-bounded by $2^{-\Lambda^\beta}$, where Λ is the codeword length. This bound holds universally for any $s/o-process$ in the family.*
3) *Both the encoding and decoding complexities are $O(|\mathcal{S}|^3 \Lambda \log \Lambda)$, where $|\mathcal{S}|$ is an upper bound on the number of states of an s/o-process in the family.*

For the proof of the theorem, see the end of Section VI.

A. Prior work on universal polar codes

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, as described below, coding methods that are based on polarization and successive-cancellation decoding 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].

B. Overview of this paper

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

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

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

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

Hochwald and Jelenković [18] considered a property similar to forgetfulness under the restrictive assumption that there is a positive probability of transitioning between any two states in one step. Leveraging ideas from Kaijser [19], who considered a related setting for hidden Markov processes, we lift this restrictive assumption and prove, in Sections IX and X, 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. 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 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. In Section VII we explain how to construct universal polar codes for a given family of processes. Numerical results for a particular universal polar code, constructed using the method of Section VII and used over several different channels with and without memory, can be found in Section VIII.

C. 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

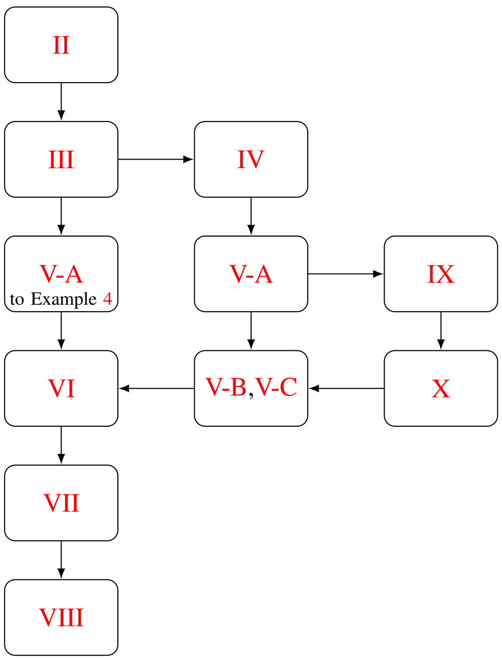

Fig. 1. Roadmap of the various ways to read this paper. All paths start at Section II and end at Section VIII.

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. The practitioner is also well-advised to read Sections VII and VIII to understand how to construct universal polar codes in practice, and to realize the benefits of using list decoding to decode these universal polar codes. We note that some definitions from Section V-C are required to follow Section VII, but we refer to the relevant equations as the need arises.
- 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 IX and X. The interested reader is advised to read them following Section V-A. Sections IX and X are written for a general hidden Markov model and may be read independently.

II. Notation and Basic Definitions

A discrete set of elements is denoted as a list in braces, e.g., $\{1, 2, \ldots, L\}$, usually denoted with a calligraphic letter, e.g., $\mathcal{A}$. The number of elements in a discrete set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. We denote $y_j^k = \begin{bmatrix} y_j & y_{j+1} & \cdots & y_k \end{bmatrix}$ 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 $X_\mathbf{b} = (X_{b_1}, X_{b_2}, \ldots, X_{b_m})$.

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

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

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

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 (η, $\mathcal{L}$, $\mathcal{H}$)-*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 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 $F_1^{N_{n+1}} = \varphi_{n+1}(U_1^{N_n}, V_1^{N_n}) = f_{n+1}(X_1^{N_{n+1}})$.

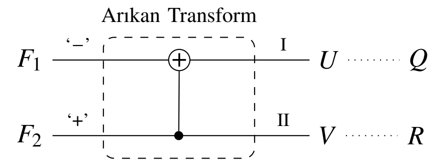

Fig. 2. Illustration of an Arıkan transform. It transforms two input symbols, U (input-I) and V (input-II) to two output symbols, F_1 (output '−') and F_2 (output '+').

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

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

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 $H(F_1|G_1) \geq \max\{H(U|Q), H(V|R)\}$ and that the inequality is strict unless either $H(U|Q)$ or $H(V|R)$ is extremal. This was the strategy of obtaining polarization for standard (Arıkan's) polar codes, with and without memory. See, for example, [2], [6], [13]. We will also pursue such a strategy.

B. Slow Polarization Stage

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

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

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

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

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 3 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 $1 \leq i \leq N_0$,

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

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

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

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

$$\begin{aligned}
F_1^{N_{n+1}} &= \varphi_{n+1}(U_1^{N_n}, V_1^{N_n}) = f_{n+1}(X_1^{N_{n+1}}), && \text{(7a)}\\
G_i &= (F_1^{i-1}, Y_1^{N_{n+1}}), \quad 1 \leq i \leq N_{n+1}. && \text{(7b)}
\end{aligned}$$

[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 $n = 0$.

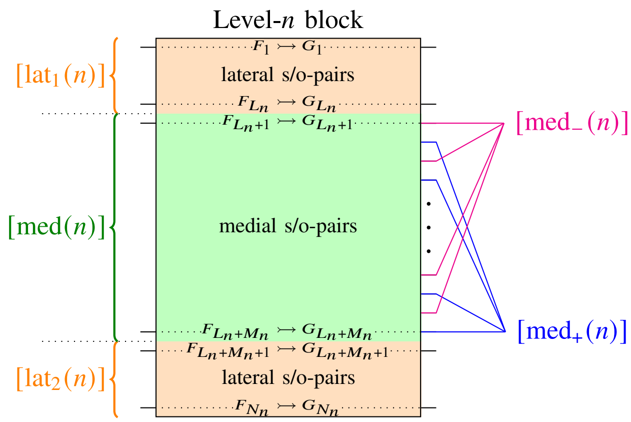

Fig. 3. Index sets in level n of the basic slow transform. A Level-n block comprises $N_n = 2L_n + M_n$ s/o-pairs. The first L_n and the last L_n s/o-pairs are lateral s/o-pairs and the remaining M_n s/o-pairs are medial s/o-pairs.

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

$$\begin{aligned}
L_{n+1} &= 2L_n + 1 && \text{(8a)}\\
M_{n+1} &= 2(M_n - 1). && \text{(8b)}
\end{aligned}$$

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

Remark 1. Observe that L_n is odd and M_n is even for any $n \geq 1$. Therefore, by (4), for any $n \geq 1$, 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 by (4) and (8),

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

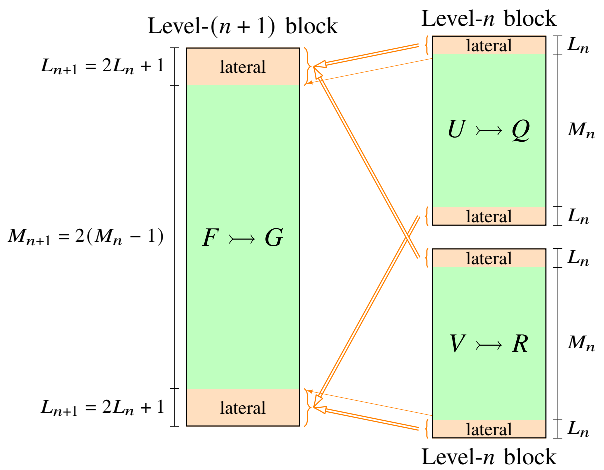

Fig. 4. A schematic description of forming lateral s/o-pairs of a level-$(n+1)$ block from two level-n blocks.

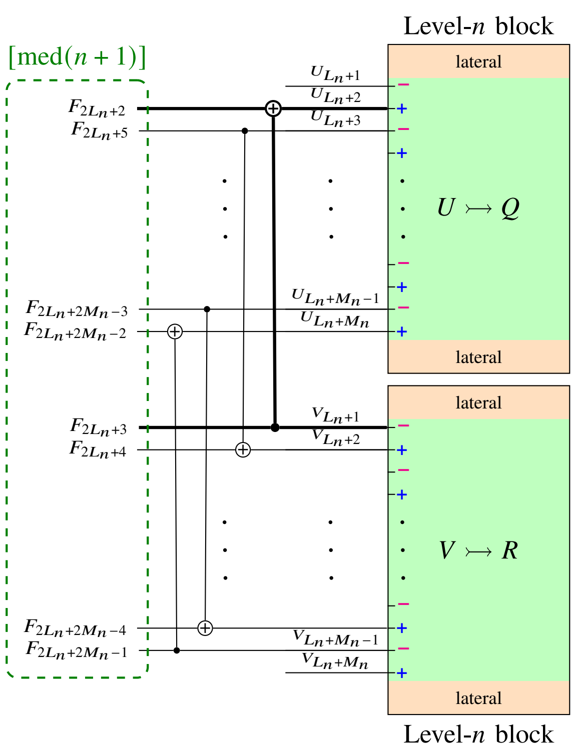

Fig. 5. Forming the medial symbols of level $n+1$ of the basic slow transform. Arıkan transforms are used with a symbol from $[\mathrm{med}_+(n)]$ of one block as their input-I and a symbol from $[\mathrm{med}_-(n)]$ of the other block as their input-II. One Arıkan transform is highlighted using thicker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

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 $n \geq 0$,

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

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).

By (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 $n \geq 1$, 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.

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 a 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 2. *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 3. *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 $n \geq 0$,

$$L_n = 2^n(L_0 + (1 - 2^{-n})) \tag{19a}$$
$$M_n = 2^n(M_0 - 2(1 - 2^{-n})). \tag{19b}$$

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 cascading multiple BSTs. This idea originates 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. For further details, see Section V-D.

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. This is also true for a cascade of BSTs, see Section V-D. Moreover, the sets $\mathcal{L}$ and $\mathcal{H}$ are predetermined and independent of the s/o-process. 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 cascade that facilitates error-free successive-cancellation decoding.

The fast polarization stage is based on Arıkan's seminal 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.

As in the proof of Theorem 1, we fix a sufficiently small η, which determines the back-off from extremality that the BST cascade will achieve. This value, as shown in Appendix A, is set small enough to ensure fast polarization of this stage. Further following the proof of Theorem 1, we set the BST cascade parameters. These include the BST parameters L_0, M_0, n, as well parameters t and $\mathbf{c}$ defining the cascade, to be discussed in Section V-D. These parameters ensure that the BST cascade in $(\eta, \mathcal{L}, \mathcal{H})$-monopolarizing and that $|\mathcal{H}|/N$ is sufficiently close to the infimal conditional entropy rate of the family of s/o-processes.

For the fast stage, take $\hat{N} = 2^{\hat{n}}$ copies of the BST cascade of length N. Apply multiple copies of Arıkan's seminal transform ("fast transform") of length $\hat{N}$, as illustrated in Figure 6. In words, the jth fast transform operates on the jth s/o-pair from each copy of the BST cascade.

As shown in the proof of Theorem 1, this construction is universal over the family of s/o-processes. That is, the code is the same for any s/o-process in the family. Recall that when decoding, we assume that the s/o-process is known at the decoder side.

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.

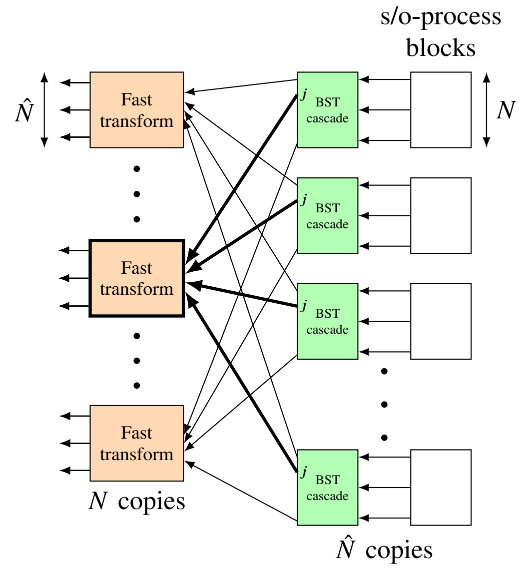

Fig. 6. Illustration of the slow and fast stages. First, $\hat{N}$ length-N blocks of the s/o-process are transformed using BST cascades of length N (denoted in green □). Then, N fast transforms of length $\hat{N}$ are applied (denoted in orange □). The jth fast transform (in bold) operates on the jth s/o-pair in each BST cascade.

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,[3] and we number them sequentially. The b-block numbered ℓ contains s/o-pairs with indices $(\ell-1)N_0+k$, $1 \le k \le N_0$. 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

$$j = (\ell - 1)N_0 + k, \quad 1 \le \ell \le 2^n, \quad 1 \le k \le N_0. \tag{20}$$

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

Recall from Definition 6 that each medial level-n symbol has 2^n medial level-0 indices as its base-ancestors. These base-ancestors are a subvector of $X_1^{N_n}$. Each of these level-0 indices has a different b-block number, computed via (20). 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.

[3]The letter 'b' here and also in the name b-index below stands for 'base,' as the BST may be thought of as consisting of 2^n "base blocks" of length N_0.

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,

$$(\mathbf{b})_\ell = (\ell - 1)N_0 + k \tag{21}$$

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

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

$$(\bar{\mathbf{b}})_\ell = (\mathbf{b})_\ell - (\ell - 1)N_0, \quad 1 \le \ell \le 2^n, \tag{22}$$

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 (21).

Remark 2. 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 (23) 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

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

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 (22). 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.

For a level-0 BST, the modulo-base-vector for medial index $L_0 + 1 \le i \le N_0 - L_0$ 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

$$\bar{\mathbf{b}}_i^{(n+1)} = \begin{bmatrix} \bar{\mathbf{b}}_{j+1}^{(n)} & \bar{\mathbf{b}}_j^{(n)} \end{bmatrix}, \quad j = \left\lfloor \frac{i}{2} \right\rfloor. \tag{23}$$

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

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

Computing a base-vector, say $\mathbf{b}_{35}^{(3)}$, is easily done using (22):

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

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

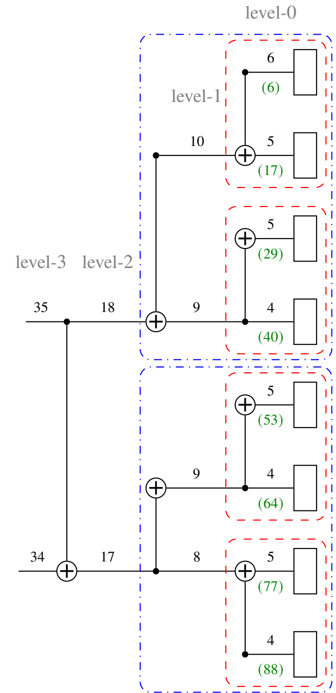

Fig. 7. A portion of a level-3 BST, initialized with $L_0 = 3, M_0 = 6$. The base-vector $\mathbf{b}_{34}^{(3)} = \mathbf{b}_{35}^{(3)}$ is illustrated. The rectangles denote level-0 BSTs. Level-1 BSTs are delimited with dashed lines (in red) and level-2 BSTs are delimited with dash-dotted lines (in 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.

Let $n \le m$. Fix some $i \in [\mathrm{med}(m)]$ and apply (23) 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:

$$\begin{aligned}\bar{\mathbf{b}}_{34}^{(3)} &= \begin{bmatrix}\begin{bmatrix}6 & 5\end{bmatrix} & \begin{bmatrix}5 & 4\end{bmatrix} & \begin{bmatrix}5 & 4\end{bmatrix} & \begin{bmatrix}5 & 4\end{bmatrix}\end{bmatrix}\\ &= \begin{bmatrix}\bar{\mathbf{b}}_{10}^{(1)} & \bar{\mathbf{b}}_{9}^{(1)} & \bar{\mathbf{b}}_{9}^{(1)} & \bar{\mathbf{b}}_{8}^{(1)}\end{bmatrix}.\end{aligned}$$

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

$$(\bar{\mathbf{b}}_{i+1}^{(n)})_\ell \ge (\bar{\mathbf{b}}_{i}^{(n)})_\ell. \tag{24}$$

Proof: This follows from (23) by straightforward induction. Specifically, note that if the index i on the left-hand-side of (23) increases, the indices j and $j+1$ on the right-hand-side cannot decrease. ■

Lemma 5. *For any $i \in [\mathrm{med}(n)]$ and any $1 \le \ell \le 2^n$ we have*

$$\left\lfloor \frac{i}{2^n} \right\rfloor = (\bar{\mathbf{b}}_{i}^{(n)})_{2^n} \le (\bar{\mathbf{b}}_{i}^{(n)})_\ell \le (\bar{\mathbf{b}}_{i}^{(n)})_1 = 1 + \left\lceil \frac{i-1}{2^n} \right\rceil. \tag{25}$$

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 (23), $\bar{\mathbf{b}}_i^{(n+1)} = \begin{bmatrix}\bar{\mathbf{b}}_{j+1}^{(n)} & \bar{\mathbf{b}}_{j}^{(n)}\end{bmatrix}$, 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 4, $(\bar{\mathbf{b}}_{j+1}^{(n)})_\ell \ge (\bar{\mathbf{b}}_{j}^{(n)})_\ell$ 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 (25): 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 (25) holds for some $n \ge 0$; we will show it holds for $n+1$ as well. Let $i \in [\mathrm{med}(n+1)]$; by (23), $(\bar{\mathbf{b}}_i^{(n+1)})_1 = (\bar{\mathbf{b}}_{j+1}^{(n)})_1$, 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,

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

where (a) is by the induction assumption and (b) is by [21, equation 3.11].

Proof of the left-hand side of (25): 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 (25) holds for some $n \ge 0$; we will show it holds for $n+1$ as well. Let $i \in [\mathrm{med}(n+1)]$; by (23), $(\bar{\mathbf{b}}_i^{(n+1)})_{2^{n+1}} = (\bar{\mathbf{b}}_j^{(n)})_{2^n}$, where $j = \lfloor i/2 \rfloor$. Therefore,

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

where (a) is by the induction assumption and (b) is by [21, equation 3.11]. ■

Corollary 6. *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 5. 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 $i \in [\mathrm{med}(0)]$:

$$f_{0,i}(X_1^{N_0}) = X_i, \tag{26a}$$
$$g_{0,i}(X_1^{N_0}, Y_1^{N_0}) = (X_1^{i-1}, Y_1^{N_0}). \tag{26b}$$

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

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

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

$$g_{n+1,i}(X_1^{N_{n+1}}, Y_1^{N_{n+1}}) = \begin{cases} \left(g_{n,j}(X_{N_n+1}^{2N_n}, Y_{N_n+1}^{2N_n}), g_{n,j+1}(X_1^{N_n}, Y_1^{N_n})\right), & i = 2j,\ j \in [\mathrm{med}_-(n)], \\ \left(g_{n,j+1}(X_1^{N_n}, Y_1^{N_n}), g_{n,j}(X_{N_n+1}^{2N_n}, Y_{N_n+1}^{2N_n})\right), & i = 2j,\ j \in [\mathrm{med}_+(n)], \\ \left(f_{n+1,i-1}(X_1^{N_{n+1}}), g_{n+1,i-1}(X_1^{N_{n+1}}, Y_1^{N_{n+1}})\right), & i = 2j+1. \end{cases} \tag{28}$$

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 13 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 (27), (28) is well-defined, as medial indices are only ever generated from medial indices (see Remark 2), 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 (27) and (28), 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 $i \in [\mathrm{med}(0)]$,

$$\tilde{f}_{0,i}(X_1^{N_0}) = X_i, \tag{29a}$$
$$\tilde{g}_{0,i}(X_1^{N_0}, Y_1^{N_0}) = (X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}). \tag{29b}$$

By comparing (26) and (29), 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 (28) 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

$$X_{\mathbf{i}} = (X_{i_1}, X_{i_2}, \ldots, X_{i_k}), \tag{30a}$$
$$X_{\mathbf{i}-L} = (X_{i_1-L}, X_{i_2-L}, \ldots, X_{i_k-L}), \tag{30b}$$
$$X_{\mathbf{i}-L}^{\mathbf{i}+M} = (X_{i_1-L}^{i_1+M}, X_{i_2-L}^{i_2+M}, \ldots, X_{i_k-L}^{i_2+M}). \tag{30c}$$

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 (26b), (28), and (29b),

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

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 (28), 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 $\mathbf{b}$, (23), the recursion (28) 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 '$\equiv$' between two vectors means that there is a one-to-one mapping between either one and the other that is independent of either vector,

$$\tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}) \equiv \left(\tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}), X_{\mathbf{b}-L_0}^{\mathbf{b}-1}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}\right). \tag{31}$$

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

$$G_i = g_{n,i}(X_{\mathbf{a}}^{\mathbf{b}}, Y_{\mathbf{a}}^{\mathbf{z}}) \equiv (\acute{G}_i, \underset{\prime}{G}_i), \tag{32}$$

where

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

This follows by induction similar to the one above. Indeed, this is obvious for the initialization step by comparing (26b) and (29b), and the induction step follows, as above, from the recursive definition of the base-vector (23) and from (28).

Remark 3. 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 j,

$$P_{S_j,X_j,Y_j|S_{-\infty}^{j-1},X_{-\infty}^{j-1},Y_{-\infty}^{j-1}} = P_{S_j,X_j,Y_j|S_{j-1}} = P_{S_j|S_{j-1}} \cdot P_{X_j,Y_j|S_j}, \tag{33}$$

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 (33) 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 probabilistic) of S_j. FAIM processes are described in detail in [14].

Remark 4. The definition of FAIM processes in [14] did not include the rightmost equality of (33). However, by suitably redefining the state of the process (for example, take (S_j, S_{j-1}) as the state at time j),[4] we may obtain the rightmost equality of (33) 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 7. *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 $L \leq M \in \mathbb{Z}$,*

$$P_{X_{-\infty}^L, Y_{-\infty}^L, X_{M+1}^\infty, Y_{M+1}^\infty} \leq \psi_{M-L} \cdot P_{X_{-\infty}^L, Y_{-\infty}^L} \cdot P_{X_{M+1}^\infty, Y_{M+1}^\infty}, \tag{34a}$$
$$P_{X_{-\infty}^L, Y_{-\infty}^L, X_{M+1}^\infty, Y_{M+1}^\infty} \geq \phi_{M-L} \cdot P_{X_{-\infty}^L, Y_{-\infty}^L} \cdot P_{X_{M+1}^\infty, Y_{M+1}^\infty}. \tag{34b}$$

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 (34a), was established in [14, Lemma 5], and the proof for (34b) is similar. For completeness, we provide a proof in Appendix B. We note at this point that for $k \geq 1$ we may take

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

in (34). 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 $k \in \mathbb{Z}$, [22, 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 7 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 7 ensures that s/o-blocks of a FAIM-derived process become almost independent when sufficiently far apart. We will need a separate property that explores what happens when a single s/o-block of a FAIM process is large enough. Specifically, we will be interested in FAIM processes that, in a sense, "forget" their past. In a forgetful FAIM process, the initial and final states of a sufficiently large s/o-block are almost independent both when given just the observations or when given the symbols and observations jointly. A precise definition of a forgetful FAIM process follows.

[4]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.

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

$$I(S_1; S_k | X_1^k, Y_1^k) \leq \epsilon, \tag{35a}$$
$$I(S_1; S_k | Y_1^k) \leq \epsilon. \tag{35b}$$

For a given λ, the infimal ϵ satisfying the above is called the *λ-forgetfulness* of the s/o-process, and is denoted ϵ_λ. Clearly, ϵ_λ is nonincreasing with λ and converges to 0. Conversely, for a given ϵ, the minimal λ is called the *ϵ-recollection* of the process.

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 X (see also Example 7 in that section). In particular, forgetful FAIM processes do exist.
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 (35a) and (35b) are required: neither condition implies the other. We demonstrate this unintuitive fact in Example 4 below.
4) In Lemma 8 below, we show that (35) together with the Markov property (33) imply that for any $k \geq \lambda$, $\ell \leq 1$, and $m \geq k$,

$$I(S_\ell; S_m | X_1^k, Y_1^k) \leq \epsilon, \tag{36a}$$
$$I(S_\ell; S_m | Y_1^k) \leq \epsilon. \tag{36b}$$

The following lemma is proved in Appendix B.

Lemma 8. *Let (S_j, X_j, Y_j), $j \in \mathbb{Z}$ be a FAIM process. Then, for any $\ell \leq 1$ and $m \geq k \geq \lambda \geq 1$, we have*

$$I(S_1; S_\lambda | X_1^\lambda, Y_1^\lambda) \geq I(S_\ell; S_m | X_1^k, Y_1^k); \tag{37a}$$
$$I(S_1; S_\lambda | Y_1^\lambda) \geq I(S_\ell; S_m | Y_1^k). \tag{37b}$$

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

$$Y_j = \begin{cases} a, & \text{if } S_j \in \{1, 2\}, \\ b, & \text{if } S_j \in \{3, 4\}. \end{cases} \tag{38}$$

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.

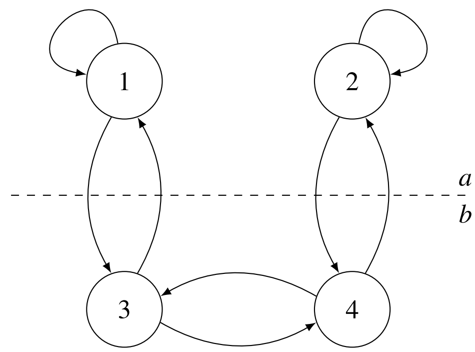

Fig. 8. The Markov chain S_j has four states. The possible transitions are depicted using arrows; the probability of choosing any transition is 1/2. The observation Y_j is 'a' if $S_j \in \{1, 2\}$ or 'b' if $S_j \in \{3, 4\}$.

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 (38). 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 (38) (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., [22, Theorem 4.3]).

Assume we have a forgetful FAIM process, and we apply to it a level-0 BST (i.e. (5)), 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 11. Moreover, we will see in Corollary 12 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 9. *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*

$$I(S_0; S_{-k}, S_k | X_{-\ell}^{-1}, Y_{-\ell}^{m}) \leq 2\epsilon. \tag{39}$$

This is a consequence of (35). 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 (33). A detailed proof can be found in Appendix B.

Lemma 9 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 10. *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 (30). 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,

$$\tilde{\mathcal{H}} \triangleq H(X_i | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0}). \tag{40}$$

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

Lemma 11. *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*

$$0 \leq \tilde{\mathcal{H}} - H(X_i | X_1^{i-1}, Y_1^{N_0}) \leq 2\epsilon. \tag{41}$$

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 (41).

Let (S_j, X_j, Y_j) be the FAIM process from which $X_\star \rightarrowtail Y_\star$ is derived. By stationarity and Lemma 9, for any m, ℓ, k such that $\min\{m, \ell\} \geq k \geq \lambda$,

$$I(S_i; S_{i-k}, S_{i+k} | X_{i-\ell}^{i-1}, Y_{i-\ell}^{i+m}) \leq 2\epsilon. \tag{42}$$

Setting $\ell = m = k = L_0$ in (42) 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 (33) 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 (33). 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 12. *Under the same setting as Lemma 11,*

$$\left|\mathcal{H}(X_\star | Y_\star) - \tilde{\mathcal{H}}\right| \leq 2\epsilon, \tag{43}$$

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 11, 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

$$(\tilde{X}_{\mathbf{b}-L_0}^{\mathbf{b}}, \tilde{Y}_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}) \sim \prod_{\ell=1}^{2^n} P_{X_{b_\ell-L_0}^{b_\ell}, Y_{b_\ell-L_0}^{b_\ell+L_0}}, \tag{44}$$

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.

Remark 5. Observe that the RHS of (44) consists of a product of distributions of "windows" of the same size. Each window, by stationarity, has the same distribution. Moreover, each window is in a different b-block. By block-independence, therefore, the RHS of (44) is independent of $\mathbf{b}$. Put another way, these windows are i.i.d. This is the crux of the results that follow: the transforms operate on an i.i.d. process.[5] In fact, the results of this section hold also for a BST operating on i.i.d. s/o-pairs. This observation will be useful in Section V-D, where we consider a cascade of BSTs, in which a step of the cascade operates on such s/o-pairs.

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

$$\tilde{F}_i = \tilde{f}_{n,i}(\tilde{X}_{\mathbf{b}}), \quad \tilde{G}_i = \tilde{g}_{n,i}(\tilde{X}_{\mathbf{b}-L_0}^{\mathbf{b}}.\tilde{Y}_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}). \tag{45a}$$
$$\tilde{F}_j = \tilde{f}_{n,j}(\tilde{X}_{\mathbf{d}}), \quad \tilde{G}_j = \tilde{g}_{n,j}(\tilde{X}_{\mathbf{d}-L_0}^{\mathbf{d}}, \tilde{Y}_{\mathbf{d}-L_0}^{\mathbf{d}+L_0}). \tag{45b}$$

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 (44) 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 (44) with $\mathbf{b}$ set to $\mathbf{d}$, the base-vector of j.

Recall from (40) that we denoted $\mathcal{H} = H(X_i | X_{i-L_0}^{i-1}, Y_{i-L_0}^{i+L_0})$, 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 $H(\tilde{F}_i|\tilde{G}_i) = \mathcal{H} + \delta_n$ and if $i \in [\mathrm{med}_+(n)]$ then $H(\tilde{F}_i|\tilde{G}_i) = \mathcal{H} - \delta_n$. This will follow as a corollary to the following lemma.

Lemma 13. *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, (29). 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 (23). Similarly, $\mathbf{d}_1$ is the base-vector of $j'+1$, and $\mathbf{d}_2$ is the base-vector of j'. Denote

$$\tilde{U}_{i'+1} = \tilde{f}_{n-1,i'+1}(\tilde{X}_{\mathbf{b}_1}), \ \tilde{Q}_{i'+1} = \tilde{g}_{n-1,i'+1}(\tilde{X}_{\mathbf{b}_1-L_0}^{\mathbf{b}_1}, \tilde{Y}_{\mathbf{b}_1-L_0}^{\mathbf{b}_1+L_0}), \tag{46a}$$
$$\tilde{V}_{i'} = \tilde{f}_{n-1,i'}(\tilde{X}_{\mathbf{b}_2}), \qquad \tilde{R}_{i'} = \tilde{g}_{n-1,i'}(\tilde{X}_{\mathbf{b}_2-L_0}^{\mathbf{b}_2}, \tilde{Y}_{\mathbf{b}_2-L_0}^{\mathbf{b}_2+L_0}). \tag{46b}$$

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 (44), $(\tilde{X}_{\mathbf{b}_1-L_0}^{\mathbf{b}_1}, \tilde{Y}_{\mathbf{b}_1-L_0}^{\mathbf{b}_1+L_0})$ 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 (27) and (28),

$$\tilde{F}_i = \tilde{U}_{i'+1} + \tilde{V}_{i'}, \quad \tilde{G}_i = \begin{cases} (\tilde{R}_{i'}, \tilde{Q}_{i'+1}), & i' \in [\mathrm{med}_-(n-1)], \\ (\tilde{Q}_{i'+1}, \tilde{R}_{i'}), & i' \in [\mathrm{med}_+(n-1)], \end{cases} \tag{47}$$

and

$$\tilde{F}_j = \tilde{U}_{j'+1} + \tilde{V}_{j'}, \quad \tilde{G}_j = \begin{cases} (\tilde{R}_{j'}, \tilde{Q}_{j'+1}), & j' \in [\mathrm{med}_-(n-1)], \\ (\tilde{Q}_{j'+1}, \tilde{R}_{j'}), & j' \in [\mathrm{med}_+(n-1)]. \end{cases} \tag{48}$$

Comparing (47) and (48), 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 $(\tilde{F}_j, \tilde{G}_j)$.

[5] In [8], the underlying process was i.i.d. to begin with; here, where there is memory, we need more intricate mechanics: the OT-BST and the BI-process.

For the case where $i, j \in [\mathrm{med}_+(n)]$, we have by (27),

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

Observe that $\tilde{F}_i$ is always a symbol in $[\mathrm{med}_-(n-1)]$. Further recall from (27) 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 (27) and (28),

$$(\tilde{F}_i, \tilde{G}_i) = (\tilde{F}_i, \tilde{F}_{i-1}, \tilde{G}_{i-1}) \equiv (\tilde{F}_i, \tilde{F}_i + \tilde{F}_{i-1}, \tilde{G}_{i-1}), \tag{49}$$

Similarly,

$$(\tilde{F}_j, \tilde{G}_j) = (\tilde{F}_j, \tilde{F}_{j-1}, \tilde{G}_{j-1}) \equiv (\tilde{F}_j, \tilde{F}_j + \tilde{F}_{j-1}, \tilde{G}_{j-1}). \tag{50}$$

The mappings on the right-hand sides of (49) and (50) are the same. Moreover, by (28), 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 14. *There exists a nondecreasing sequence $\delta_n \geq 0$, 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 $H(\tilde{F}_{i+1}|\tilde{G}_{i+1}) = \tilde{\mathcal{H}} - \delta_n$.*

Observe from (4d) and (4e) that Corollary 14 implies that there exists a nondecreasing sequence $\delta_n \geq 0$ such that

$$H(\tilde{F}_i|\tilde{G}_i) = \begin{cases} \tilde{\mathcal{H}} + \delta_n, & i \in [\mathrm{med}_-(n)], \\ \tilde{\mathcal{H}} - \delta_n, & i \in [\mathrm{med}_+(n)]. \end{cases} \tag{51}$$

Further observe that Corollary 14 implies that for any $i \in [\mathrm{med}_-(n)]$ and $j \in [\mathrm{med}_+(n)]$ we have

$$H(\tilde{F}_i|\tilde{G}_i) + H(\tilde{F}_j|\tilde{G}_j) = 2\tilde{\mathcal{H}}. \tag{52}$$

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 $i \in [\mathrm{med}_-(n)]$ 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 (45a) and let $\tilde{U}_{i'+1}, \tilde{V}_{i'}, \tilde{Q}_{i'+1}, \tilde{R}_{i'}$ be defined as in (46). We have, by (27) and (28),

$$\begin{aligned} H(\tilde{F}_i|\tilde{G}_i) + H(\tilde{F}_{i+1}|\tilde{G}_{i+1}) &= H(\tilde{F}_i, \tilde{F}_{i+1}|\tilde{Q}_{i'+1}, \tilde{R}_{i'}) \\ &= H(\tilde{U}_{i'+1}, \tilde{V}_{i'}|\tilde{Q}_{i'+1}, \tilde{R}_{i'}) \\ &= H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1}) + H(\tilde{V}_{i'}|\tilde{R}_{i'}), \end{aligned} \tag{53}$$

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

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

Thus,

$$H(\tilde{F}_i|\tilde{G}_i) + H(\tilde{F}_{i+1}|\tilde{G}_{i+1}) = 2\tilde{\mathcal{H}}. \tag{54}$$

By (4d) and (4e) and since $i \in [\mathrm{med}_-(n)]$, we have $i+1 \in [\mathrm{med}_+(n)]$. Recall from Remark 1 that since $n \geq 1$ then i is even and $i+1$ is odd. By (27), (28), and since conditioning reduces entropy, we have

$$\begin{aligned} H(\tilde{F}_{i+1}|\tilde{G}_{i+1}) &\leq \min\{H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1}), H(\tilde{V}_{i'+1}|\tilde{R}_{i'+1})\} \\ &= \tilde{\mathcal{H}} - \delta_{n-1}. \end{aligned} \tag{55}$$

From (54) and (55), 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 13, 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 14 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 16 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

$$\alpha * \beta = \alpha + \beta(1-2\alpha) = \beta + \alpha(1-2\beta), \tag{56}$$

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 15. *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 $\xi_1, \xi_2 > 0$,*

$$\sum_{a=1}^k p_a h_2(\alpha_a) \geq \xi_1, \qquad \sum_{b=1}^k q_b h_2(\beta_b) \leq \xi_2, \tag{57}$$

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 $i \in [\mathrm{med}(n)]$, 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 $i \in [\mathrm{med}_-(n)]$, so that i is even, and $i' = i/2$. We define $\tilde{F}_i, \tilde{G}_i$ as in (45a), and $\tilde{U}_{i'+1}, \tilde{V}_{i'}, \tilde{Q}_{i'+1}, \tilde{R}_{i'}$ as in (46).

Lemma 16. *For all $\xi > 0$, if $i \in [\mathrm{med}_-(n)]$ and*

$$H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1}), H(\tilde{V}_{i'}|\tilde{R}_{i'}) \in (\xi, 1-\xi) \tag{58}$$

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

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

Recall that $\tilde{U}_{i'+1}, \tilde{V}_{i'}$ are binary, so the minimizations are between two terms. As a result, $0 \le \alpha_q, \beta_r \le 1/2$. With this notation and by (58) we have

$$H(\tilde{U}_{i'+1}|\tilde{Q}_{i'+1}) = \sum_q \tilde{p}(q) h_2(\alpha_q) \ge \xi,$$
$$H(\tilde{V}_{i'}|\tilde{R}_{i'}) = \sum_r \tilde{p}(r) h_2(\beta_r) \le 1 - \xi.$$

Thus, by (47) and the independence of $(\tilde{U}_{i'+1}, \tilde{Q}_{i'+1})$ and $(\tilde{V}_{i'}, \tilde{R}_{i'})$, we obtain

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

where the inequality is by Lemma 15. ■

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 (40).

Proposition 17. *For every $\xi > 0$, there exists a threshold value $n_{\text{th}} \ge 0$ such that if $n \ge n_{\text{th}}$ then a level-n OT-BST with any parameters L_0, M_0 is $(\xi, [\text{med}_+(n)], [\text{med}_-(n)])$-monopolarizing for any 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 $n \ge n_{\text{th}}$. Then:

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

Proof: We first consider a specific selection of the transform parameters L_0, M_0, and a specific BI-process $\tilde{X}_\star \rightarrowtail \tilde{Y}_\star$ with parameter $N_0 = 2L_0 + M_0$. For these selections we will find a threshold n'_{th}. We will then find a global upper bound n_{th} on n'_{th}, which is independent of these selections.

Denote the indicator functions

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

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

Observe by Corollary 14 that if $\mathcal{M}_n^- = 1$ then also $H(\tilde{F}_i|\tilde{G}_i) < 2\tilde{\mathcal{H}} - (1-\xi)$, and if $\mathcal{M}_n^+ = 1$ then also $H(\tilde{F}_i|\tilde{G}_i) > 2\tilde{\mathcal{H}} - \xi$.

Further denote

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

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. Moreover, by Corollary 14, $\mathcal{M}_n = 1$ if and only if the bulleted items in the claim hold.

We define

$$n'_{\text{th}} = \min\left\{n \in \mathbb{N} \mid \mathcal{M}_n = 1\right\},$$

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

By Corollary 14, there exists a nondecreasing sequence $\delta_n \ge 0$ such that (51) holds. Since δ_n is a nondecreasing sequence, $\mathcal{M}_n = 1$ for every $n \ge n'_{\text{th}}$. The entropy of a binary random variable is bounded between 0 and 1; thus for any n, $0 \le \tilde{\mathcal{H}} - \delta_n \le \tilde{\mathcal{H}} + \delta_n \le 1$. Hence, $\delta_n \le \min\{\tilde{\mathcal{H}}, 1 - \tilde{\mathcal{H}}\}$. We conclude that if $\tilde{\mathcal{H}} \le 1/2$ and $n \ge n'_{\text{th}}$ then $\mathcal{M}_n^+ = 1$, and if $\tilde{\mathcal{H}} \le 1/2$ and $n \ge n'_{\text{th}}$ 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 $n \ge 0$, $\mathcal{M}_n = 0$, then by (51) and by definition of $\mathcal{M}_n^-$, $\mathcal{M}_n^+$, we obtain

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

Rearranging, this yields

$$\mathcal{M}_n = 0 \Rightarrow \delta_n \le \min\{\tilde{\mathcal{H}}, 1 - \tilde{\mathcal{H}}\} - \xi. \tag{59}$$

On the other hand, by (51) and Lemma 16, if $\mathcal{M}_{n-1} = 0$ for some $n \ge 1$, we have

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

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

$$\mathcal{M}_{n-1} = 0 \Rightarrow \delta_n \ge n\Delta(\xi, 1-\xi). \tag{60}$$

Now, let

$$n_1 = 1 + \left\lfloor \frac{\min\{\tilde{\mathcal{H}}, 1 - \tilde{\mathcal{H}}\} - \xi}{\Delta(\xi, 1-\xi)} \right\rfloor, \tag{61}$$

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 (59) and (60) we obtain

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

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

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

which contradicts (61) (see, e.g., [21, Equation 3.3]). We conclude that we must have $n'_{\text{th}} \le n_1$.

We have found an upper bound for n'_{th}, which is given by the RHS of (61). Note that this bound is indeed positive, thus it holds for both cases of $\mathcal{M}_0$ discussed above. Next, observe that

$$1 + \left\lfloor \frac{\min\{\tilde{\mathcal{H}}, 1 - \tilde{\mathcal{H}}\} - \xi}{\Delta(\xi, 1-\xi)} \right\rfloor \le 1 + \left\lfloor \frac{1/2 - \xi}{\Delta(\xi, 1-\xi)} \right\rfloor.$$

Thus, defining

$$n_{\text{th}} = 1 + \left\lfloor \frac{1/2 - \xi}{\Delta(\xi, 1-\xi)} \right\rfloor \tag{62}$$

suffices. ■

Remark 6. Note that if $\tilde{\mathcal{H}}$ is given to us, then n_{th} can be taken as the RHS of (61).

The following corollary is a restatement of Proposition 17 that will be useful in the sequel for proving Proposition 29.

Corollary 18. *Let $\tilde{\mathcal{H}}$, $H(\tilde{F}_i|\tilde{G}_i)$, ξ and $n \geq n_{\text{th}}$ be as in Proposition 17. Define*

$$\begin{aligned}\alpha(\tilde{\mathcal{H}}) &= \min\{2\tilde{\mathcal{H}}, 1\},\\ \alpha'(\tilde{\mathcal{H}}) &= \alpha(\tilde{\mathcal{H}}) - \xi,\\ \beta(\tilde{\mathcal{H}}) &= \max\{2\tilde{\mathcal{H}} - 1, 0\},\\ \beta'(\tilde{\mathcal{H}}) &= \beta(\tilde{\mathcal{H}}) + \xi.\end{aligned}$$

Then,

$$H(\tilde{F}_i|\tilde{G}_i) \in \begin{cases} \Big(\alpha'(\tilde{\mathcal{H}}), \alpha(\tilde{\mathcal{H}})\Big] & i \in [\text{med}_-(n)] \quad \text{(63a)}\\ \Big[\beta(\tilde{\mathcal{H}}), \beta'(\tilde{\mathcal{H}})\Big) & i \in [\text{med}_+(n)]. \quad \text{(63b)}\end{cases}$$

Proof: Observe that $\alpha(\tilde{\mathcal{H}}) + \beta(\tilde{\mathcal{H}}) = 2\tilde{\mathcal{H}}$, by considering separately the cases $\tilde{\mathcal{H}} \leq 1/2$ and $\tilde{\mathcal{H}} \geq 1/2$. Clearly, $\alpha'(\tilde{\mathcal{H}}) + \beta'(\tilde{\mathcal{H}}) = 2\tilde{\mathcal{H}}$ as well. This implies, by (52), that if (63a) holds for some $\tilde{\mathcal{H}}$ then (63b) also holds, and vice versa.

Consider first the case $\tilde{\mathcal{H}} \leq 1/2$ and let $i \in [\text{med}_+(n)]$. In this case

$$\beta(\tilde{\mathcal{H}}) = 0, \ \beta'(\tilde{\mathcal{H}}) = \xi.$$

By Proposition 17, $H(\tilde{F}_i|\tilde{G}_i) \in [\beta(\tilde{\mathcal{H}}), \beta'(\tilde{\mathcal{H}}))$. Hence, in this case (63b) holds, and by the discussion above (63a) must also hold for this case.

Next, consider the case $\tilde{\mathcal{H}} \geq 1/2$ and let $i \in [\text{med}_-(n)]$. In this case

$$\alpha(\tilde{\mathcal{H}}) = 1, \ \alpha'(\tilde{\mathcal{H}}) = 1 - \xi.$$

By Proposition 17, $H(\tilde{F}_i|\tilde{G}_i) \in (\alpha'(\tilde{\mathcal{H}}), \alpha(\tilde{\mathcal{H}})]$. Hence, in this case (63a) holds, and by the discussion above (63b) must also hold for this case. ■

The following corollarly will be used in the proof of Theorem 21.

Corollary 19. *For a given $\xi > 0$, let L_0, M_0, and n_{th} be as in Proposition 17. Then, under the same setting as Proposition 17, 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 $H(\tilde{F}_i|\tilde{G}_i) > 1 - \xi - \zeta$, $\forall i \in [\text{med}_-(n)]$.*

Proof: This corollary follows from Proposition 17 and Corollary 14. Recall that by Corollary 14, there exists $\delta_n \geq 0$ such that (51) 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 17. Otherwise, $\tilde{\mathcal{H}} \geq 1/2$, so by Proposition 17 and (51),

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

Rearranging, we obtain $\delta_n > 1 - \tilde{\mathcal{H}} - \xi$. Now, by (51),

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

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 17 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

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

Lemma 20. *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 $i \in [\text{med}_+(n)]$ then

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

Proof: In light of Corollary 14, denote, for any $n \geq 0$ and arbitrary $i \in [\text{med}_-(n)]$

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

Observe that for arbitrary $i \in [\text{med}_+(n)]$, by Corollary 14 we have $\delta_n = \tilde{\mathcal{H}} - H(\tilde{F}_i|\tilde{G}_i)$. Our goal is thus to show that for any $n \geq 0$,

$$C_n(\tilde{\mathcal{H}}) \leq \delta_n \leq D_n(\tilde{\mathcal{H}}). \tag{64}$$

The remainder of the proof mirrors the proof of [8, Proposition 2]. We prove the claim by induction. If $n = 0$, the claim is trivially true. Assume that the claim holds for some $n \geq 0$, and we will show it also holds for $n+1$.

By block-independence of the BI-process we may use [6, Lemma 2.1], by which

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

Rearranging, we obtain

$$c(\delta_n, \tilde{\mathcal{H}}) \leq \delta_{n+1} \leq d(\delta_n, \tilde{\mathcal{H}}). \tag{65}$$

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

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

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 $r(\cdot) \leq 1/2$ by definition. Thus, by (65) and the induction hypothesis (64),

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

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 $(0.004, [\text{med}_+(n)], [\text{med}_-(n)])$-monopolarizing for the BI-process whenever $n \geq n_{\text{th}}$.

Proposition 17, even when using the tighter bound in Remark 6, gives the upper bound

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

This is a prohibitive value. Thankfully, it is also unnecessarily pessimistic. To obtain a practical value for n_{th}, we turn to Lemma 20, by which

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

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

C. Monopolarization for FAIM-derived Processes, for a Single BST

This subsection is devoted to proving Theorem 21 below. That is, we show that the BST is monopolarizing for suitably chosen η, $\mathcal{L}$, $\mathcal{H}$ when applied to a set of forgetful[6] FAIM-derived s/o-processes. The theorem holds for a set of s/o-processes that satisfies a set of rather lax conditions. The conditions are defined in terms of the L-forgetfulness of the s/o-process, ϵ_L, see Definition 10; and the mixing parameters ψ_M and ϕ_M from Lemma 7.

Recall that if a *specific* FAIM-derived s/o-process satisfies Condition K, it is forgetful (see Example 7). Thus, by Definition 10, we have that $\epsilon_L \xrightarrow[L\to\infty]{} 0$. Furthermore, for this FAIM-derived s/o-process, by Lemma 7, $\psi_M \xrightarrow[M\to\infty]{} 1$, and $\phi_M \xrightarrow[M\to\infty]{} 1$, where ψ_M is nonincreasing and ϕ_M is nondecreasing. Our conditions for the set of forgetful FAIM-derived s/o-processes are that there exist sequences $\bar{\epsilon}_L$, $\bar{\psi}_M$, and $\bar{\phi}_M$ such that for *any* s/o-process in the set we have

$$\epsilon_L \leq \bar{\epsilon}_L \xrightarrow[L\to\infty]{} 0, \tag{66a}$$

$$\psi_M \leq \bar{\psi}_M \xrightarrow[M\to\infty]{} 1, \tag{66b}$$

$$\phi_M \geq \bar{\phi}_M \xrightarrow[M\to\infty]{} 1, \tag{66c}$$

$$\bar{\psi}_M \text{ nonincreasing}, \quad \bar{\phi}_M \text{ nondecreasing}. \tag{66d}$$

Clearly, for a given finite set $\mathcal{C}$ of forgetful FAIM-derived s/o-processes, conditions (66) are easy to satisfy. Namely, for any L and M, take $\bar{\epsilon}_L = \max_{c\in\mathcal{C}} \epsilon_L(c)$, $\bar{\psi}_M = \max_{c\in\mathcal{C}} \psi_M(c)$, and $\bar{\phi}_M = \min_{c\in\mathcal{C}} \phi_M(c)$. For an infinite set $\mathcal{C}$, we may replace max and min above by sup and inf, respectively.

Theorem 21. *Fix sequences $\bar{\epsilon}_L$, $\bar{\psi}_M$, and $\bar{\phi}_M$ that satisfy the limits in (66a)–(66c), as well as the conditions in (66d). Let $X_\star \rightarrowtail Y_\star$ be a forgetful FAIM-derived s/o-process that satisfies the inequalities in (66a)–(66c). For every $\eta > 0$ there exist L_{th}, M_{th}, and n_{th},* independent of the process, *such that if $L_0 \geq L_{\text{th}}$, $M_0 \geq M_{\text{th}}$, and $n \geq n_{\text{th}}$ then a level-n BST initialized with parameters L_0 and M_0 is $(\eta, [\text{med}_+(n)], [\text{med}_-(n)])$-monopolarizing.*

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

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

This theorem will follow as a corollary to Proposition 22 below. We will show in Proposition 22 that, when L_{th} and M_{th} 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 17, the OT-BST of a BI-process is monopolarizing, this will imply that the BST is also monopolarizing.

The parameters L_{th}, M_{th}, and n_{th} for given sequences $\bar{\epsilon}_L$, $\bar{\psi}_M$, and $\bar{\phi}_M$ are determined in the proof of Theorem 21. For future reference, they are detailed in the following remark.

Remark 7. We are given $\eta > 0$, and our goal is to set the parameters n_{th}, L_{th}, and M_{th}. We do this via the following steps.

1) Fix $\varepsilon_1 < \eta/12$ and $\varepsilon_2 < \eta^2/32$, in line with (85) below.
2) Let ξ be as in (84) below.
3) Set n_{th} using this ξ in (62).
4) Take $\epsilon = \varepsilon_1 2^{-n_{\text{th}}}$, in line with (77).
5) Find L_{th} such that (35) holds with this ϵ, and L_{th} in place of k.
6) Find M_{th} such that (80) holds with M_{th} in place of M, and n_{th} in place of n.

These steps will follow from the proof of Theorem 21.

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

[6] An interesting open problem is whether the forgetfulness condition is necessary.

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

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

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

$$F_i = f_{n,i}(X_{\mathbf{b}}), \quad G_i = g_{n,i}(X_{\mathbf{a}}^{\mathbf{b}}, Y_{\mathbf{a}}^{\mathbf{z}}), \tag{67a}$$
$$\acute{F}_i = \tilde{f}_{n,i}(X_{\mathbf{b}}), \quad \acute{G}_i = \tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}), \tag{67b}$$
$$\tilde{F}_i = \tilde{f}_{n,i}(\tilde{X}_{\mathbf{b}}), \quad \tilde{G}_i = \tilde{g}_{n,i}(\tilde{X}_{\mathbf{b}-L_0}^{\mathbf{b}}, \tilde{Y}_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}). \tag{67c}$$

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.

The following proposition, as well as Lemmas 23 and 24, are stated for a forgetful FAIM-derived s/o-process, $X_\star \rightarrowtail Y_\star$, that satisfies (66) for some sequences $\bar{\epsilon}_L$, $\bar{\psi}_M$, and $\bar{\phi}_M$.

Proposition 22. *Fix $n \geq 0$, $\varepsilon_1 > 0$, and $0 < \varepsilon_2 < \frac{1}{6}$. There exist L_{th} and M_{th} such that a level-n BST initialized with parameters $L_0 \geq L_{\mathrm{th}}, M_0 \geq M_{\mathrm{th}}$ satisfies:*

$$|H(F_i|G_i) - H(\tilde{F}_i|\tilde{G}_i)| \leq 2\varepsilon_1 + \frac{\varepsilon_2}{2} - 3\varepsilon_2 \log \frac{3\varepsilon_2}{2} \tag{68a}$$
$$< 2\varepsilon_1 + \sqrt{8\varepsilon_2}. \tag{68b}$$

Furthermore, we have

$$|H(F_i|G_i, S_0, S_{N_n}) - H(\tilde{F}_i|\tilde{G}_i)| \leq 2\varepsilon_1 + \frac{\varepsilon_2}{2} - 3\varepsilon_2 \log \frac{3\varepsilon_2}{2}. \tag{69}$$

Proof: Denote

$$\acute{P} = P_{X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}},$$
$$\tilde{P} = \prod_{\ell=1}^{2^n} P_{X_{b_\ell-L_0}^{b_\ell}, Y_{b_\ell-L_0}^{b_\ell+L_0}}.$$

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

In Lemma 23 that follows we show that there exists L_{th} such that if $L_0 \geq L_{\mathrm{th}}$ then

$$|H(F_i|G_i) - H(\acute{F}_i|\acute{G}_i)| \leq 2\varepsilon_1. \tag{70}$$

Next, in Lemma 24 that follows we show that there exists M_{th} such that if $M_0 \geq M_{\mathrm{th}}$ then

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

This will enable us to use Lemma 25 below with $f = \tilde{f}_{n,i}$ and $g = \tilde{g}_{n,i}$ to obtain

$$|H(\acute{F}_i|\acute{G}_i) - H(\tilde{F}_i|\tilde{G}_i)| \leq \frac{\varepsilon_2}{2} - 3\varepsilon_2 \log \frac{3\varepsilon_2}{2} < \sqrt{8\varepsilon_2}.$$

Hence, we conclude that

$$\begin{aligned} &|H(F_i|G_i) - H(\tilde{F}_i|\tilde{G}_i)| \\ &\quad \leq |H(F_i|G_i) - H(\acute{F}_i|\acute{G}_i)| + |H(\acute{F}_i|\acute{G}_i) - H(\tilde{F}_i|\tilde{G}_i)| \\ &\quad \leq 2\varepsilon_1 + \frac{\varepsilon_2}{2} - 3\varepsilon_2 \log \frac{3\varepsilon_2}{2} \\ &\quad < 2\varepsilon_1 + \sqrt{8\varepsilon_2}. \end{aligned}$$

This proves (68).

Finally, to show (69), by (73) in Lemma 23 we have

$$|H(F_i|G_i, S_0, S_{N_n}) - H(\acute{F}_i|\acute{G}_i)| \leq 2\varepsilon_1.$$

We insert the above in place of (70), and by the same arguments as before,

$$\begin{aligned} &|H(F_i|G_i, S_0, S_{N_n}) - H(\tilde{F}_i|\tilde{G}_i)| \\ &\quad \leq |H(F_i|G_i, S_0, S_{N_n}) - H(\acute{F}_i|\acute{G}_i)| + |H(\acute{F}_i|\acute{G}_i) - H(\tilde{F}_i|\tilde{G}_i)| \\ &\quad \leq 2\varepsilon_1 + \frac{\varepsilon_2}{2} - 3\varepsilon_2 \log \frac{3\varepsilon_2}{2}. \end{aligned}$$

This completes the proof. ∎

In the sequel, we will refer to the right-hand side of (68a). To this end, for $\varepsilon_1 > 0$ and $0 < \varepsilon_2 < \frac{1}{6}$, we denote

$$\varepsilon_3 = \varepsilon_3(\varepsilon_1, \varepsilon_2) \triangleq 2\varepsilon_1 + \frac{\varepsilon_2}{2} - 3\varepsilon_2 \log \frac{3\varepsilon_2}{2}. \tag{71}$$

We now state and prove Lemmas 23 to 25.

For the lemma below, recall that a BST is initialized with parameters L_0 and M_0. This lemma is concerned with L_0 and applies for any M_0.

Lemma 23. *Fix $n \geq 0$ and $\varepsilon_1 > 0$. There exists L_{th} such that if $L_0 \geq L_{\mathrm{th}}$ then for all M_0,*

$$0 \leq H(\acute{F}_i|\acute{G}_i) - H(F_i|G_i) \leq 2\varepsilon_1 \tag{72}$$

and

$$0 \leq H(\acute{F}_i|\acute{G}_i) - H(F_i|G_i, S_0, S_{N_n}) \leq 2\varepsilon_1. \tag{73}$$

Proof: By (32), $G_i \equiv (\acute{G}_i, \underset{\smile}{G}_i)$, where

$$\acute{G}_i = \tilde{g}_{n,i}(X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}),$$
$$\underset{\smile}{G}_i = (X_{\mathbf{a}}^{\mathbf{b}-L_0-1}, Y_{\mathbf{a}}^{\mathbf{b}-L_0-1}, Y_{\mathbf{b}+L_0+1}^{\mathbf{z}}). \tag{74}$$

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

$$H(F_i|G_i) = H(\acute{F}_i|\acute{G}_i, \underset{\smile}{G}_i) \leq H(\acute{F}_i|\acute{G}_i), \tag{75}$$

where the inequality is because conditioning reduces entropy. This proves the left-hand side of (72).

We now turn to proving the right-hand side of (72). Utilizing the left-hand side of (66a) and Corollary 10 with $\mathbf{i} = \mathbf{b}$, $\lambda = L_{\mathrm{th}}$, and $L_0 \geq L_{\mathrm{th}}$, we obtain

$$I(S_{\mathbf{b}}; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid X_{\mathbf{b}-L_0}^{\mathbf{b}-1}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}) \leq 2^n \cdot 2\bar{\epsilon}_L. \tag{76}$$

Using the right-hand side of (66a), we take L_{th} large enough so that

$$\bar{\epsilon}_{L_{\mathrm{th}}} \leq \varepsilon_1 \cdot 2^{-n}. \tag{77}$$

Hence,

$$
\begin{aligned}
2\varepsilon_1 &\geq I(S_{\mathbf{b}}; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \\
&\overset{\text{(a)}}{\geq} I(\acute{F}_i, \acute{G}_i; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \\
&\overset{\text{(b)}}{\geq} I(\acute{F}_i; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid \acute{G}_i, X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \\
&\overset{\text{(c)}}{=} I(\acute{F}_i; S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0} \mid \acute{G}_i) \\
&= H(\acute{F}_i | \acute{G}_i) - H(\acute{F}_i | \acute{G}_i, S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) \\
&\overset{\text{(d)}}{=} H(\acute{F}_i | \acute{G}_i) - H(\acute{F}_i | \acute{G}_i, \underaccent{\prime}{G}_i, S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) \\
&\overset{\text{(e)}}{=} H(\acute{F}_i | \acute{G}_i) - H(F_i | G_i, S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) \qquad (78) \\
&\overset{\text{(f)}}{\geq} H(\acute{F}_i | \acute{G}_i) - H(F_i | G_i),
\end{aligned}
$$

where:

- (a) is due to (2). By (33), $X_{\mathbf{b}}$ is a probabilistic function of $S_{\mathbf{b}}$; by (67b), $\acute{F}_i$ 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

$$
\begin{aligned}
(S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) &\multimap (S_{\mathbf{b}}, X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \\
&\multimap (X_{\mathbf{b}}, X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \multimap (\acute{F}_i, \acute{G}_i).
\end{aligned}
$$

  Specifically, we have the Markov chain

$$
(S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) \multimap (S_{\mathbf{b}}, X^{\mathbf{b}-1}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}) \multimap (\acute{F}_i, \acute{G}_i),
$$

  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 (31) and (67b).
- (d) is by the Markov property (33), (67b), and (74): $\acute{F}_i$ and $\acute{G}_i$ are probabilistic functions of states $S^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}$, whereas $\underaccent{\prime}{G}_i$ 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 $G_i \equiv (\acute{G}_i, \underaccent{\prime}{G}_i)$ by (32).
- (f) is because conditioning reduces entropy.

This shows (72).

Finally, to show (73) we use the same steps. Namely, we obtain the left-hand side of (73) by replacing (75) with

$$
H(F_i | G_i, S_0, S_{N_n}) = H(\acute{F}_i | \acute{G}_i, \underaccent{\prime}{G}_i, S_0, S_{N_n}) \leq H(\acute{F}_i | \acute{G}_i).
$$

For the right-hand side of (73), observe that by the penultimate step of (78) we have

$$
\begin{aligned}
2\varepsilon_1 &\geq H(\acute{F}_i | \acute{G}_i) - H(F_i | G_i, S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}) \\
&= H(\acute{F}_i | \acute{G}_i) - H(F_i | G_i, S_{\mathbf{b}-L_0}, S_{\mathbf{b}+L_0}, S_0, S_{N_n}) \\
&\geq H(\acute{F}_i | \acute{G}_i) - H(F_i | G_i, S_0, S_{N_n}).
\end{aligned}
$$

The first equality is by Markov property (33). The final inequality is because conditioning reduces entropy.

Lastly, note that there are no restrictions on M_0 throughout the proof — its only role is setting the parameters of the BST — and thus the claim holds for any M_0. ■

For the lemma below, again recall that a BST is initialized with parameters L_0 and M_0. This lemma is concerned with M_0 and applies for any L_0.

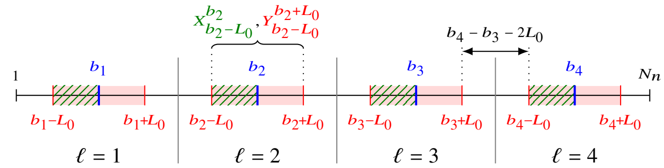

Fig. 9. Illustration of a level-2 BST. There are four b-blocks, with b-block numbers $\ell = 1, 2, 3, 4$. The base-index (in blue) in b-block ℓ is b_ℓ. The red boxes in the illustration correspond to $X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}$, where X is only available to 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.

Lemma 24. *Fix $n \geq 0$, $\varepsilon_2 > 0$, and L_0. There exists M_{th} such that if $M_0 \geq M_{\text{th}}$ then*

$$
P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} \leq (1+\varepsilon_2) \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}, \qquad (79a)
$$

$$
P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} \geq (1-\varepsilon_2) \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}. \qquad (79b)
$$

Proof: By the right-hand sides of (66b) and (66c), we may choose M_{th} such that

$$
(\bar{\psi}_{M_{\text{th}}-2})^{(2^n)} \leq 1 + \varepsilon_2, \qquad (80a)
$$

$$
(\bar{\phi}_{M_{\text{th}}-2})^{(2^n)} \geq 1 - \varepsilon_2. \qquad (80b)
$$

For any $M_0 \geq M_{\text{th}}$, by (66d), we thus have

$$
(\bar{\psi}_{M_0-2})^{(2^n)} \leq (\bar{\psi}_{M_{\text{th}}-2})^{(2^n)} \leq 1 + \varepsilon_2, \qquad (81a)
$$

$$
(\bar{\phi}_{M_0-2})^{(2^n)} \geq (\bar{\phi}_{M_{\text{th}}-2})^{(2^n)} \geq 1 - \varepsilon_2. \qquad (81b)
$$

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 6, for any $1 \leq \ell < 2^n$ we have $1 \leq |\bar{b}_{\ell+1} - \bar{b}_\ell| \leq 2$. Hence, by (22), and recalling that $N_0 = 2L_0 + M_0$,

$$
\begin{aligned}
(b_{\ell+1} - L_0) - (b_\ell + L_0) &= \ell N_0 - (\ell-1)N_0 - 2L_0 + \bar{b}_{\ell+1} - \bar{b}_\ell \\
&= M_0 + (\bar{b}_{\ell+1} - \bar{b}_\ell) \\
&\geq M_0 - |\bar{b}_{\ell+1} - \bar{b}_\ell| \\
&\geq M_0 - 2. \qquad (82)
\end{aligned}
$$

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 $\mathcal{B}_\ell = \{b_\ell - L_0, b_\ell - L_0 + 1, \ldots, b_\ell + L_0\}$, $1 \leq \ell \leq 2^n$. 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 (82), any two consecutive sets $\mathcal{B}_\ell$ and $\mathcal{B}_{\ell+1}$ are separated by at least $M_0 - 2$ indices.

Using the left-hand sides of (66b) and (66c), Lemma 7, and a straightforward induction argument, we conclude that

$$
P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} \leq (\bar{\psi}_{M_0-2})^{(2^n)} \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}},
$$

$$
P_{X^{\mathbf{b}}_{\mathbf{b}-L_0}, Y^{\mathbf{b}+L_0}_{\mathbf{b}-L_0}} \geq (\bar{\phi}_{M_0-2})^{(2^n)} \prod_{\ell=1}^{2^n} P_{X^{b_\ell}_{b_\ell-L_0}, Y^{b_\ell+L_0}_{b_\ell-L_0}}.
$$

Thus, by (81),

$$P_{X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}} \leq (1+\varepsilon_2) \prod_{\ell=1}^{2^n} P_{X_{b_\ell - L_0}^{b_\ell}, Y_{b_\ell - L_0}^{b_\ell + L_0}},$$
$$P_{X_{\mathbf{b}-L_0}^{\mathbf{b}}, Y_{\mathbf{b}-L_0}^{\mathbf{b}+L_0}} \geq (1-\varepsilon_2) \prod_{\ell=1}^{2^n} P_{X_{b_\ell - L_0}^{b_\ell}, Y_{b_\ell - L_0}^{b_\ell + L_0}},$$

which is (79). ■

In Lemma 24 we saw that $\acute{P}$ and $\tilde{P}$ are close in the sense of (79). The following lemma, whose proof can be found in Appendix D, translates this proximity to conditional entropies.

Lemma 25. *Let A and $\tilde{A}$ be two discrete random variables over the same finite alphabet $\mathcal{A}$. Denote $\mathbb{P}(A=a) = p(a)$ and $\mathbb{P}(\tilde{A}=a) = q(a)$ for all $a \in \mathcal{A}$. Assume that for some $0 \leq \varepsilon < \frac{1}{6}$,*

$$(1-\varepsilon)q(a) \leq p(a) \leq (1+\varepsilon)q(a), \quad \forall a \in \mathcal{A}. \tag{83}$$

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

$$\left|H(f(A)|g(A)) - H(f(\tilde{A})|g(\tilde{A}))\right| \leq \frac{\varepsilon}{2} - 3\varepsilon \log \frac{3\varepsilon}{2} < \sqrt{8\varepsilon}.$$

We are now ready to prove Theorem 21.

Proof of Theorem 21: Choose $\varepsilon_1 > 0$ and $0 < \varepsilon_2 < \frac{1}{6}$ small enough such that

$$\xi \triangleq \eta - 4\varepsilon_1 - (2\varepsilon_1 + \sqrt{8\varepsilon_2}) > 0. \tag{84}$$

For example, one may take

$$\varepsilon_1 < \frac{\eta}{12}, \tag{85a}$$
$$\varepsilon_2 < \frac{\eta^2}{32}. \tag{85b}$$

Take n_{th} large enough so that Proposition 17 holds with ξ as above. Recall that Proposition 17 holds for any L_0 and M_0, so we are free to set them as desired.

By Proposition 22, for n_{th}, ε_1, and ε_2 above, there exist L_{th} and M_{th} such that (68) holds for $L_0 \geq L_{\text{th}}$ and $M_0 \geq M_{\text{th}}$. That is,

$$-(2\varepsilon_1 + \sqrt{8\varepsilon_2}) \leq H(F_i|G_i) - H(\tilde{F}_i|\tilde{G}_i) \leq (2\varepsilon_1 + \sqrt{8\varepsilon_2}). \tag{86}$$

In fact, we choose $L_0 \geq L_{\text{th}}$ as in the proof of Lemma 23. This ensures that the L_0-forgetfulness of the s/o-process is upper-bounded by ε_1. Thus, by Corollary 12, (43) holds with $\epsilon \leq \varepsilon_1$, so that

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

Hence, if $\mathcal{H}(X_\star|Y_\star) \leq 1/2$ then $\tilde{\mathcal{H}} \leq (1+4\varepsilon_1)/2$ and if $\mathcal{H}(X_\star|Y_\star) \geq 1/2$ then $\tilde{\mathcal{H}} \geq (1-4\varepsilon_1)/2$. Consequently, by Corollary 19 with $\zeta = 4\varepsilon_1$, if $n \geq n_{\text{th}}$ then

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

Combining the above with (84) and (86) we obtain that for $n \geq n_{\text{th}}$,

$$\mathcal{H}(X_\star|Y_\star) \leq 1/2 \Rightarrow H(F_i|G_i) < \eta, \quad \forall i \in [\text{med}_+(n)],$$
$$\mathcal{H}(X_\star|Y_\star) \geq 1/2 \Rightarrow H(F_i|G_i) > 1-\eta, \quad \forall i \in [\text{med}_-(n)].$$

This completes the proof. ■

Remark 8. We have proved Theorem 21 using (68b), which is looser than (68a). Hence, the proof would also hold if we were to define ξ in (84) as $\eta - 4\varepsilon_1 - \varepsilon_3$. The looser definition of ξ circumvents a cumbersome upper bound on ε_3, and by proxy on ε_2, that involves the Lambert W function [23, p. 332].

D. Monopolarization for FAIM-derived Processes, for a Cascade of BSTs

The previous subsection considered a single BST. However, in practice, one may cascade several BSTs to obtain a universal polar code of rate different from 1/2. In this subsection, we extend the previous results to a cascade of BSTs.

A cascade of BSTs is defined by the following:

- the number of BSTs in the cascade, t (that is, the overall transform is comprised of t stages);
- parameter L_0, which defines the number of lateral indices in the level-0 block of the first BST in the cascade;
- parameters $M_0^{\{1\}}, M_0^{\{2\}}, \ldots, M_0^{\{t\}}$, where $M_0^{\{i\}}$ defines the number of medial indices in the level-0 block of the ith BST in the cascade;
- recursion depths $n_1, n_2, \ldots, n_t$, of the t BSTs in the cascade;
- a binary vector $\mathbf{c} = \begin{bmatrix} c_1 & c_2 & \ldots & c_{t-1} \end{bmatrix}$ of length $t-1$. Each stage of the cascade applies a BST operation on a subset of indices from the previous stage. This is determined by $\mathbf{c}$; informally, $c_i = 0$ ($c_i = 1$) implies that stage $i+1$ is the result of applying a BST on the newly formed medial-minus (plus) indices of stage i.

The cascade is constructed recursively. For $t = 1$, we are in the single BST case. This BST is defined through L_0, parameter $M_0 = M_0^{\{1\}}$, and recursion depth n_1. That is, we transform $X_1^{N^{\{1\}}}$ to $U_1^{N^{\{1\}}}$, where $N^{\{1\}} = (2L_0 + M_0^{\{1\}}) \cdot 2^{n_1}$. Recall that medial indices at level n_1 of the BST are split into $[\text{med}_-(n_1)]$ and $[\text{med}_+(n_1)]$, see (4). We denote the medial sets of stage 1 of the cascade as

$$[\text{med}_-\{1\}] = [\text{med}_-(n_1)], \quad [\text{med}_+\{1\}] = [\text{med}_+(n_1)].$$

The remaining indices are lateral. We further define the two sets $[\nu\text{med}_-\{1\}]$ and $[\nu\text{med}_+\{1\}]$, which we call “the new medial-minus and medial-plus sets of stage 1 of the cascade.” For this base case, they coincide with $[\text{med}_-\{1\}]$ and $[\text{med}_+\{1\}]$, respectively. That is,

$$[\nu\text{med}_-\{1\}] = [\text{med}_-\{1\}], \quad [\nu\text{med}_+\{1\}] = [\text{med}_+\{1\}].$$

When moving from stage i to stage $i+1$ of the cascade, we first make $M_0^{\{i+1\}} \cdot 2^{n_{i+1}}$ copies of the stage-i cascade. That is, the length of the stage-$(i+1)$ cascade is

$$N^{\{i+1\}} = N^{\{i\}} \cdot M_0^{\{i+1\}} \cdot 2^{n_{i+1}}. \tag{87}$$

For $1 \le \ell \le M_0^{\{i+1\}} \cdot 2^{n_{i+1}}$ and $1 \le j \le N^{\{i\}}$, Denote

$$X_j^{(\ell)} = X_{j+(\ell-1)N^{\{i\}}},$$
$$U_j^{(\ell)} = U_{j+(\ell-1)N^{\{i\}}},$$

and their vector versions

$$\boldsymbol{X}_\ell = \begin{bmatrix} X_1^{(\ell)} & X_2^{(\ell)} & \cdots & X_{N^{\{i\}}}^{(\ell)} \end{bmatrix},$$
$$\boldsymbol{U}_\ell = \begin{bmatrix} U_1^{(\ell)} & U_2^{(\ell)} & \cdots & U_{N^{\{i\}}}^{(\ell)} \end{bmatrix}.$$

Copy ℓ of the stage-i cascade transforms $\boldsymbol{X}_\ell$ to $\boldsymbol{U}_\ell$. Next, for each $1 \le j \le N^{(i)}$, we take the jth index of each copy and apply, as described below, an operation: either a BST or a pass-through. That is, for $1 \le j \le N^{\{i\}}$ and $1 \le \ell \le M_0^{\{i+1\}} \cdot 2^{n_{i+1}}$, further denote

$$V_\ell^{(j)} = V_{\ell+(j-1)\cdot M_0^{\{i+1\}}\cdot 2^{n_{i+1}}}, \tag{88}$$

and the vectors

$$\underline{\boldsymbol{U}}_j = \begin{bmatrix} U_j^{(1)} & U_j^{(2)} & \cdots & U_j^{(M_0^{\{i+1\}}\cdot 2^{n_{i+1}})} \end{bmatrix}, \tag{89a}$$
$$\underline{\boldsymbol{V}}_j = \begin{bmatrix} V_1^{(j)} & V_2^{(j)} & \cdots & V_{M_0^{\{i+1\}}\cdot 2^{n_{i+1}}}^{(j)} \end{bmatrix}. \tag{89b}$$

The operation transforms $\underline{\boldsymbol{U}}_j$ to $\underline{\boldsymbol{V}}_j$. The output of the stage-$(i+1)$ cascade is $\underline{\boldsymbol{V}}_1, \underline{\boldsymbol{V}}_2, \ldots, \underline{\boldsymbol{V}}_{N^{\{i\}}}$. In other words, we operate on a single symbol from each copy, and the result is a single contiguous block on the output side. This ordering is amenable to successive-cancellation decoding.

What remains to define is which operation, BST or pass-through, to apply to which index, and to determine the various medial sets. At stage i of the cascade, each index is either medial or lateral. The medial indices are split into $[\text{med}_-\{i\}]$ and $[\text{med}_+\{i\}]$. These sets will be defined as part of the recursion. Important subsets of these sets are $[\nu\text{med}_-\{i\}]$ and $[\nu\text{med}_+\{i\}]$, respectively. These two subsets will also be defined as part of the recursion. BST operations will be applied to exactly one of these subsets, called the "active set." On all other indices, the pass-through operation will be applied.

When moving from stage $i < t$ to stage $i+1$, define the "active set" σ_i as

$$\sigma_i = \begin{cases} [\nu\text{med}_-\{i\}], & \text{if } c_i = 0, \\ [\nu\text{med}_+\{i\}], & \text{if } c_i = 1. \end{cases} \tag{90}$$

For $i = t$ we technically define the active set as the descendants of the active set of the previous stage, that is

$$\sigma_t = [\nu\text{med}_-\{t\}] \cup [\nu\text{med}_+\{t\}]. \tag{91}$$

For each index $1 \le j \le N^{\{i\}}$, we do the following:

- If $j \in \sigma_i$, apply a BST operation to $\underline{\boldsymbol{U}}_j$. The BST is defined by parameters $L_0 = 0$, $M_0 = M_0^{\{i+1\}}$, and has recursion depth n_{i+1}.
  - If $V_\ell^{(j)}$ is a medial-minus symbol, then (see (88))
  $$\ell + (j-1)\cdot M_0^{\{i+1\}} \cdot 2^{n_{i+1}} \in [\nu\text{med}_-\{i+1\}].$$
  - If $V_\ell^{(j)}$ is a medial-plus symbol, then
  $$\ell + (j-1)\cdot M_0^{\{i+1\}} \cdot 2^{n_{i+1}} \in [\nu\text{med}_+\{i+1\}].$$
  - All indices in $[\nu\text{med}_-\{i+1\}]$ are also in $[\text{med}_-\{i+1\}]$.
  - All indices in $[\nu\text{med}_+\{i+1\}]$ are also in $[\text{med}_+\{i+1\}]$.
- Otherwise, apply a pass-through operation to $\underline{\boldsymbol{U}}_j$, that is $\underline{\boldsymbol{V}}_j = \underline{\boldsymbol{U}}_j$, i.e., by (89), $V_\ell^{(j)} = U_j^{(\ell)}$ for each $1 \le \ell \le M_0^{\{i+1\}} \cdot 2^{n_{i+1}}$.
  - If $j \in [\text{med}_-\{i\}]$, then all the indices in $\underline{\boldsymbol{V}}_j$ (see (88) and (89b)), are in $[\text{med}_-\{i+1\}]$.
  - If $j \in [\text{med}_+\{i\}]$, then all the indices in $\underline{\boldsymbol{V}}_j$ are in $[\text{med}_+\{i+1\}]$.
- Any index that is not in $[\text{med}_-\{i\}]$ or $[\text{med}_+\{i\}]$ is lateral.

We introduce the following definition — a specialization of the above — to simplify the statements of the claims in this subsection.

Definition 12. A $(t, \mathbf{c}; L_0, M_0, n)$-cascade is a cascade of BSTs as above, with $M_0^{\{i\}} = M_0$ and $n_i = n$, for $1 \le i \le t$.

The length of a $(t, \mathbf{c}; L_0, M_0, n)$-cascade, by the recursion (87) and recalling that $N^{\{1\}} = (2L_0 + M_0)2^n$ is given by

$$N^{\{t\}} = (2L_0 + M_0) M_0^{t-1} 2^{nt}. \tag{92}$$

Our plan for the rest of this subsection is as follows. Denote the cascade threshold entropy as

$$h(\mathbf{c}) = \frac{1 + \sum_{i=1}^{t-1} c_i 2^{t-i}}{2^t}. \tag{93}$$

In Lemma 26, we show that the fraction of medial-minus indices out of all medial indices of the cascade approaches $h(\mathbf{c})$. We further show that the fraction of medial indices out of all indices approaches 1. Then, in Theorem 28, we show that $h(\mathbf{c})$ is indeed a threshold entropy of the cascade. That is, for an s/o-process with conditional entropy rate less than $h(\mathbf{c})$, the medial-plus indices monopolarize; and for an s/o-process with conditional entropy rate greater than $h(\mathbf{c})$, the medial-minus indices monopolarize.

Lemma 26. *Consider a $(t, \mathbf{c}; L_0, M_0, n)$-cascade. Then,*

$$\frac{|[\text{med}_-\{t\}]|}{|[\text{med}_+\{t\}] \cup [\text{med}_-\{t\}]|} = \frac{1 + \sum_{i=1}^{t-1} c_i 2^{t-i} \left(\frac{1}{1 - 2(1-2^{-n})M_0^{-1}}\right)^{t-i}}{2 + \sum_{i=1}^{t-1} 2^{t-i} \left(\frac{1}{1 - 2(1-2^{-n})M_0^{-1}}\right)^{t-i}}. \tag{94}$$

Moreover,

$$\frac{|[\text{med}_+\{t\}] \cup [\text{med}_-\{t\}]|}{N^{\{t\}}} = \frac{\left(1 - \frac{2 - 2^{1-n}}{M_0}\right)^t \left(2 + \sum_{i=1}^{t-1} 2^{t-i} \left(\frac{1}{1 - 2(1-2^{-n})M_0^{-1}}\right)^{t-i}\right)}{2^t (2L_0 + M_0) M_0^{-1}}. \tag{95}$$

Observe that when M_0 is large, the right-hand side of (94) approaches $h(\mathbf{c})$, as the terms in parentheses approach 1 and

$\sum_{i=1}^{t-1} 2^{t-i} = 2^t - 2$. Furthermore, when M_0 is also large with respect to L_0, the right-hand side of (95) approaches 1.

Proof: Define $\nu_0 = 1$, and denote by ν_i the number of new medial plus indices (which is the same as the number of new medial minus indices) at the output of stage i of the cascade. By the cascade construction, the first stage is merely a BST of depth n with b-blocks consisting of M_0 medial indices and $2L_0$ lateral indices each. Recalling that the number of medial indices in a BST of depth n is given in (19b), we have

$$\nu_1 = \frac{1}{2}\nu_0 \cdot (2^n M_0 - 2(2^n - 1))\,.$$

Note that the factor $1/2$ stems from counting the number of new medial plus indices, which is half the number of new medial indices.

When moving from step $i-1$ of the cascade to step i, we perform ν_{i-1} BSTs of depth n, each with b-blocks consisting of M_0 medial indices and 0 lateral indices. Hence, the number of new medial plus indices at the end of stage i is

$$\nu_i = \frac{1}{2}\nu_{i-1} \cdot (2^n M_0 - 2(2^n - 1))\,. \tag{96}$$

Thus,

$$\nu_i = \frac{1}{2^i}\left(2^n M_0 - 2(2^n - 1)\right)^i\,. \tag{97}$$

We now claim that for $1 \le i \le t$,

$$|[\mathrm{med}_-\{i\}]| = \nu_i + \sum_{j=1}^{i-1} c_j \nu_j (2^n M_0)^{i-j} \tag{98a}$$

$$|[\mathrm{med}_+\{i\}]| = \nu_i + \sum_{j=1}^{i-1} (1-c_j) \nu_j (2^n M_0)^{i-j}. \tag{98b}$$

Indeed, by the construction above, if $c_i = 0$,

$$|[\mathrm{med}_-\{i+1\}]| = \nu_{i+1} + 2^n M_0 |[\mathrm{med}_-\{i\}] \setminus [\nu\mathrm{med}_-\{i\}]|,$$
$$|[\mathrm{med}_+\{i+1\}]| = \nu_{i+1} + 2^n M_0 |[\mathrm{med}_+\{i\}]|,$$

and if $c_i = 1$ then

$$|[\mathrm{med}_-\{i+1\}]| = \nu_{i+1} + 2^n M_0 |[\mathrm{med}_-\{i\}]|,$$
$$|[\mathrm{med}_+\{i+1\}]| = \nu_{i+1} + 2^n M_0 |[\mathrm{med}_+\{i\}] \setminus [\nu\mathrm{med}_+\{i\}]|.$$

Recalling that $|[\nu\mathrm{med}_-\{i\}]| = |[\nu\mathrm{med}_+\{i\}]| = \nu_i$, we can use the above to prove (98) by induction on i.

Next, observe that

$$\begin{aligned} |[\mathrm{med}_-\{i\}] \cup [\mathrm{med}_+\{i\}]| &= |[\mathrm{med}_-\{i\}]| + |[\mathrm{med}_+\{i\}]| \\ &= 2\nu_i + \sum_{j=1}^{i-1} \nu_j (2^n M_0)^{i-j}. \end{aligned} \tag{99}$$

Denote $\alpha = 2(2^n - 1)$. By (97) and (98a),

$$\begin{aligned} |[\mathrm{med}_-\{t\}]| &= \nu_t + \sum_{i=1}^{t-1} c_i \nu_i (2^n M_0)^{t-i} \\ &= \frac{(2^n M_0 - \alpha)^t}{2^t} + \sum_{i=1}^{t-1} c_i 2^{-i} (2^n M_0 - \alpha)^i (2^n M_0)^{t-i} \\ &= \frac{(2^n M_0 - \alpha)^t}{2^t} + \frac{(2^n M_0 - \alpha)^t}{2^t} \sum_{i=1}^{t-1} c_i 2^{t-i} \left(\frac{2^n M_0}{2^n M_0 - \alpha}\right)^{t-i} \\ &= \frac{(2^n M_0 - \alpha)^t}{2^t} \left(1 + \sum_{i=1}^{t-1} c_i 2^{t-i} \left(\frac{1}{1 - \alpha 2^{-n} M_0^{-1}}\right)^{t-i}\right). \end{aligned}$$

Similarly, by (97) and (99),

$$\begin{aligned} |[\mathrm{med}_-\{t\}] \cup [\mathrm{med}_+\{t\}]| &= 2\nu_t + \sum_{i=1}^{t-1} \nu_i (2^n M_0)^{t-i} \\ &= 2 \cdot \frac{(2^n M_0 - \alpha)^t}{2^t} + \sum_{i=1}^{t-1} 2^{-i} (2^n M_0 - \alpha)^i (2^n M_0)^{t-i} \\ &= 2 \cdot \frac{(2^n M_0 - \alpha)^t}{2^t} + \frac{(2^n M_0 - \alpha)^t}{2^t} \left(\sum_{i=1}^{t-1} 2^{t-i} \left(\frac{2^n M_0}{2^n M_0 - \alpha}\right)^{t-i}\right) \\ &= \frac{(2^n M_0 - \alpha)^t}{2^t} \left(2 + \sum_{i=1}^{t-1} 2^{t-i} \left(\frac{1}{1 - \alpha 2^{-n} M_0^{-1}}\right)^{t-i}\right). \end{aligned} \tag{100}$$

Combining the above two expressions, we obtain

$$\begin{aligned} &\frac{|[\mathrm{med}_-\{t\}]|}{|[\mathrm{med}_+\{t\}] \cup [\mathrm{med}_-\{t\}]|} \\ &= \frac{1 + \sum_{i=1}^{t-1} c_i 2^{t-i} \left(\frac{1}{1 - \alpha 2^{-n} M_0^{-1}}\right)^{t-i}}{2 + \sum_{i=1}^{t-1} 2^{t-i} \left(\frac{1}{1 - \alpha 2^{-n} M_0^{-1}}\right)^{t-i}}, \\ &= \frac{1 + \sum_{i=1}^{t-1} c_i 2^{t-i} \left(\frac{1}{1 - 2(1-2^{-n}) M_0^{-1}}\right)^{t-i}}{2 + \sum_{i=1}^{t-1} 2^{t-i} \left(\frac{1}{1 - 2(1-2^{-n}) M_0^{-1}}\right)^{t-i}}, \end{aligned}$$

where in the last equality we recalled that $\alpha = 2(2^n - 1)$. This proves (94).

To prove (95), we divide the expression in (100) by the expression in (92). Since $\alpha = 2(2^n - 1)$, the ratio between the term preceding the large parentheses in (100) and $N^{\{t\}}$ is

$$\begin{aligned} \frac{(2^n M_0 - \alpha)^t}{2^t N^{\{t\}}} &= \frac{(2^n (M_0 - 2) + 2)^t}{2^t (2L_0 + M_0) M_0^{t-1} 2^{nt}} \\ &= \frac{(M_0 - 2 + 2^{1-n})^t}{2^t (2L_0 + M_0) M_0^{t-1}} \\ &= \frac{(1 - (2 - 2^{1-n}) M_0^{-1})^t}{2^t (2L_0 + M_0) M_0^{-1}}. \end{aligned}$$

This completes the proof. ∎

The following simple observations on $h(\mathbf{c})$ will be useful in the proof of Proposition 29 below.

Lemma 27. *Let* $\mathbf{c}$ *be a binary vector of length* $t-1$ *and let its length-*$(t-2)$ *suffix obtained by removing* c_1 *be* $\mathbf{c}^1$ *Then:*

$$h(\mathbf{c}^1) = 2h(\mathbf{c}) - c_1, \tag{101}$$

$$\frac{1}{2^{t-1}} \le h(\mathbf{c}^1) \le 1 - \frac{1}{2^{t-1}}. \tag{102}$$

Observe that by (101) and (102) we have $h(\mathbf{c}) \le 1/2 - 2^{-t}$ if $c_1 = 0$ and $h(\mathbf{c}) \ge 1/2 + 2^{-t}$ if $c_1 = 1$.

Proof: The proof follows from simple algebra. Indeed,

$$\begin{aligned} h(\mathbf{c}^1) &= \frac{1 + \sum_{i=1}^{t-2} c_{i+1} 2^{t-1-i}}{2^{t-1}} = 2 \cdot \left(\frac{1 + \sum_{i=1}^{t-1} c_i 2^{t-i}}{2^t} - \frac{c_1}{2}\right) \\ &= 2h(\mathbf{c}) - c_1, \end{aligned}$$

which is (101). Next, by setting $c_{i+1} = 0$ ($c_{i+1} = 1$) for all $1 \le i \le t-2$ in the expression for $h(\mathbf{c}^1)$, i.e., the first equality above, we obtain the lower (upper) bound in (102). ■

In the sequel we will denote by $\mathbf{c}^\ell$ the suffix of $\mathbf{c}$ after removing its first ℓ elements. That is,

$$\mathbf{c}^\ell = \begin{bmatrix} c_{\ell+1} & c_{\ell+2} & \cdots & c_{t-1} \end{bmatrix}. \tag{103}$$

Note that $\mathbf{c}^{t-1}$ is an empty vector, and by (93) we have $h(\mathbf{c}^{t-1}) = 1/2$.

The following is the cascade equivalent of Theorem 21.

Theorem 28. *Let the number of stages of the cascade t, and the binary vector $\mathbf{c}$ of length $t-1$ be given. Fix sequences $\bar{\epsilon}_L$, $\bar{\psi}_M$, and $\bar{\phi}_M$ that satisfy the limits in (66a)–(66c), as well as the conditions in (66d). Let $X_\star \rightarrowtail Y_\star$ be a forgetful FAIM-derived s/o-process that satisfies the inequalities in (66a)–(66c). For every $\eta > 0$ there exist L_{th}, M_{th}, and n_{th},* independent of the process, *such that if $L_0 \ge L_{\text{th}}$, $M_0 \ge M_{\text{th}}$, and $n \ge n_{\text{th}}$ then a $(t, \mathbf{c}; L_0, M_0, n)$-cascade is $(\eta, [\text{med}_+\{t\}], [\text{med}_-\{t\}])$-monopolarizing.*

Specifically, let $F_1^{N^{\{t\}}} \rightarrowtail G_1^{N^{\{t\}}}$ be a transformed s/o-block of a $(t, \mathbf{c}; L_0, M_0, n)$-cascade as in Definition 12. Then:

- *if $\mathcal{H}(X_\star|Y_\star) \le h(\mathbf{c})$ then $H(F_i|G_i) < \eta$, $\forall i \in [\text{med}_+\{t\}]$;*
- *if $\mathcal{H}(X_\star|Y_\star) \ge h(\mathbf{c})$ then $H(F_i|G_i) > 1-\eta$, $\forall i \in [\text{med}_-\{t\}]$.*

The proof will follow along the same general lines of Theorem 21. We first consider the simple case of an observation-truncated transform applied to a BI-process. For this simple case, we generalize Proposition 17.

Recall that in a BI-process with parameter $N_0 = 2L_0 + M_0$, contiguous symbol and observation blocks of length N_0 are independent. In our setting, a BI-process is defined as in Definition 11, with the transform length N_n replaced by $N^{\{t\}}$ and the number of copies of the level-0 block of length N_0 (2^n in the definition) is $N^{\{t\}}/N_0$. Further recall that an observation-truncated transform is defined in Section IV-B. The key point to note is that the "observation-truncated" property is defined based on a truncation at a level-0 block of a BST, see (29). In other words, this property is determined at the "input-output" level of the process. Consequently, in a cascade of more than one BST, the overall transform is observation truncated if the first BST is. We call such a cascade an observation-truncated cascade.

The following is a generalization of Proposition 17. Recall that $\tilde{\mathcal{H}}$ was defined in (40), and that σ_t was defined in (91).

Proposition 29. *Fix cascade parameters t and $\mathbf{c}$. For every $\zeta > 0$, there exists a threshold value $n_{\text{th}} \ge 0$ such that if $n \ge n_{\text{th}}$ then an observation-truncated $(t, \mathbf{c}; L_0, M_0, n)$-cascade with any parameters L_0, M_0 is $(\zeta, [\text{med}_+\{t\}], [\text{med}_-\{t\}])$-monopolarizing for any BI-process $\tilde{X}_\star \rightarrowtail \tilde{Y}_\star$ with parameter $N_0 = 2L_0 + M_0$.*

Specifically, let $\tilde{F}_1^{N^{\{t\}}} \rightarrowtail \tilde{G}_1^{N^{\{t\}}}$ be an OT-transformed s/o-block of the observation-truncated $(t, \mathbf{c}; L_0, M_0, n)$-cascade, where $n \ge n_{\text{th}}$. Then:

$$\tilde{\mathcal{H}} \le h(\mathbf{c}) \Rightarrow H(\tilde{F}_i|\tilde{G}_i) < \zeta, \qquad \forall i \in [\text{med}_+\{t\}], \tag{104a}$$
$$\tilde{\mathcal{H}} \ge h(\mathbf{c}) \Rightarrow H(\tilde{F}_i|\tilde{G}_i) > 1-\zeta, \quad \forall i \in [\text{med}_-\{t\}]. \tag{104b}$$

In fact, we can strengthen the above:

$$\tilde{\mathcal{H}} \le h(\mathbf{c}) \Rightarrow H(\tilde{F}_i|\tilde{G}_i) < \frac{\zeta}{2^t}, \qquad \forall i \in [\text{med}_+\{t\}] \setminus \sigma_t, \tag{105a}$$
$$\tilde{\mathcal{H}} \ge h(\mathbf{c}) \Rightarrow H(\tilde{F}_i|\tilde{G}_i) > 1-\frac{\zeta}{2^t}, \quad \forall i \in [\text{med}_-\{t\}] \setminus \sigma_t. \tag{105b}$$

Proof: First observe that if the proposition holds for some $\zeta < 1$, then it clearly holds for all $\zeta \ge 1$ with the same n_{th}. Thus, assume that $\zeta < 1$.

We start by choosing n_{th} such that Proposition 17 holds with $\xi = \zeta/2^t$. First consider the case $\tilde{\mathcal{H}} = h(\mathbf{c})$; the more general case will follow by monotonicity. We now track the evolution of the cascade stages, by using Corollary 18.

If $c_1 = 0$, then, by Lemma 27, $h(\mathbf{c}) \le 1/2$. Hence, by Corollary 18 and (101),

$$H(\tilde{F}_i|\tilde{G}_i) \in \begin{cases} (h(\mathbf{c}^1) - \xi, h(\mathbf{c}^1)], & i \in [\text{med}_-\{1\}], \\ [0, \xi), & i \in [\text{med}_+\{1\}]. \end{cases}$$

Similarly, if $c_1 = 1$, then, by Lemma 27, $h(\mathbf{c}) \ge 1/2$. Hence, by Corollary 18 and (101),

$$H(\tilde{F}_i|\tilde{G}_i) \in \begin{cases} (1-\xi, 1], & i \in [\text{med}_-\{1\}], \\ [h(\mathbf{c}^1), h(\mathbf{c}^1) + \xi), & i \in [\text{med}_+\{1\}]. \end{cases}$$

Recall that by the cascade construction, $[\nu\text{med}_-\{1\}] = [\text{med}_-\{1\}]$ and $[\nu\text{med}_+\{1\}] = [\text{med}_+\{1\}]$. Thus, the active set σ_1 (see (90)) is not polarized, whereas the remaining new medial set is polarized in the sense of (105). In particular, regardless of whether $c_1 = 0$ or $c_1 = 1$, we have

$$i \in \sigma_1 \Longrightarrow H(\tilde{F}_i|\tilde{G}_i) \in (h(\mathbf{c}^1) - \xi, h(\mathbf{c}^1) + \xi).$$

This will form the basis of the following claim, which we prove by induction: after $\ell < t$ stages of the cascade,

$$H(\tilde{F}_i|\tilde{G}_i) \in \begin{cases} (h(\mathbf{c}^\ell) - (2^\ell - 1)\xi, h(\mathbf{c}^\ell) + (2^\ell - 1)\xi), & i \in \sigma_\ell, \\ (1-\xi, 1], & i \in [\text{med}_-\{\ell\}] \setminus \sigma_\ell, \\ [0, \xi), & i \in [\text{med}_+\{\ell\}] \setminus \sigma_\ell. \end{cases}$$

The claim implies that after $\ell < t$ stages of the cascade, all the medial indices are polarized in the sense of (105), except for the active set.

As shown above, the claim is indeed true for the basis case, $\ell = 1$. For the induction step, assume the claim is true after $\ell < t-1$ stages. To prove that it is also true after $\ell + 1$ stages, we call upon Corollary 18 and Remark 5. That is, recall that to form stage $\ell + 1$, we make $M_0^{\{\ell+1\}} \cdot 2^{n_{\ell+1}}$ copies of a symbol in the active set, and apply a BST to the copies. Further recall that we are in a BI-process setting. That is, these copies are i.i.d. By Remark 5 and the induction hypothesis, Corollary 18 holds with $\tilde{\mathcal{H}}$ replaced by some value

$$\eta \in (h(\mathbf{c}^\ell) - (2^\ell - 1)\xi, h(\mathbf{c}^\ell) + (2^\ell - 1)\xi). \tag{106}$$

For η as in (106), if $c_{\ell+1} = 0$, then, by Lemma 27, $h(\mathbf{c}^\ell) \le 1/2 - 2^{-(t-\ell)}$. By our assumption that $\zeta < 1$, we have $\xi = \zeta/2^t < 2^{-t}$. Hence, by (106),

$$\eta < h(\mathbf{c}^\ell) + (2^\ell - 1)\xi \le \frac{1}{2} + 2^\ell \cdot (\xi - 2^{-t}) - \xi < \frac{1}{2}.$$

Therefore, by Corollary 18,

$$H(\tilde{F}_i|\tilde{G}_i) \in \begin{cases} (2\eta-\xi, 2\eta], & i \in [\nu\mathrm{med}_-\{\ell+1\}], \\ [0,\xi), & i \in [\nu\mathrm{med}_+\{\ell+1\}]. \end{cases} \tag{107}$$

Using (101) and (106), equation (107) implies

$$\begin{aligned} &i \in [\nu\mathrm{med}_-\{\ell+1\}] \Longrightarrow \\ &H(\tilde{F}_i|\tilde{G}_i) \in (h(\mathbf{c}^{\ell+1}) - (2^{\ell+1}-1)\xi, h(\mathbf{c}^{\ell+1}) + (2^{\ell+1}-2)\xi). \end{aligned}$$

Similarly, for η as in (106), if $c_{\ell+1} = 1$, then $\eta > 1/2$, so by Corollary 18,

$$H(\tilde{F}_i|\tilde{G}_i) \in \begin{cases} (1-\xi, 1], & i \in [\nu\mathrm{med}_-\{\ell+1\}], \\ [2\eta-1, 2\eta-1+\xi), & i \in [\nu\mathrm{med}_+\{\ell+1\}]. \end{cases}$$

Again, Using (101) and (106), we obtain that in this case

$$\begin{aligned} &i \in [\nu\mathrm{med}_+\{\ell+1\}] \Longrightarrow \\ &H(\tilde{F}_i|\tilde{G}_i) \in (h(\mathbf{c}^{\ell+1}) - (2^{\ell+1}-2)\xi, h(\mathbf{c}^{\ell+1}) + (2^{\ell+1}-1)\xi). \end{aligned}$$

In other words, combining both of the above cases and recalling the definition of the active set (90), we have

$$\begin{aligned} &i \in \sigma_{\ell+1} \Longrightarrow \\ &H(\tilde{F}_i|\tilde{G}_i) \in (h(\mathbf{c}^{\ell+1}) - (2^{\ell+1}-1)\xi, h(\mathbf{c}^{\ell+1}) + (2^{\ell+1}-1)\xi). \end{aligned}$$

Recalling the cascade construction, and specifically how $[\mathrm{med}_-\{\ell+1\}]$ and $[\mathrm{med}_+\{\ell+1\}]$ are obtained from $[\mathrm{med}_-\{\ell\}]$, $[\mathrm{med}_+\{\ell\}]$, $[\nu\mathrm{med}_-\{\ell+1\}]$, and $[\nu\mathrm{med}_+\{\ell+1\}]$, we obtain the remainder of the inductive claim. Using (91), this proves (105).

For the last stage of the cascade, t, we call upon Corollary 18. By the cascade construction, we need only consider the active set σ_{t-1}, as all other indices are polarized (in the sense of (105) and thus also in the sense of (104)). Recall that $h(\mathbf{c}^{t-1}) = 1/2$, and thus, by the inductive claim, and since $\zeta = \xi \cdot 2^t$,

$$i \in \sigma_{t-1} \Longrightarrow H(\tilde{F}_i|\tilde{G}_i) \in \left(\frac{1}{2} - \frac{\zeta}{2} + \xi, \frac{1}{2} + \frac{\zeta}{2} - \xi\right).$$

By (63a) and the monotonicity of $\alpha(\cdot)$ and $\alpha'(\cdot)$,

$$\begin{aligned} i \in [\nu\mathrm{med}_-\{t\}] \Longrightarrow H(\tilde{F}_i|\tilde{G}_i) &> \alpha'(1/2 - \zeta/2 + \xi) \\ &= 1 - \zeta + 2\xi - \xi \\ &> 1 - \zeta. \end{aligned}$$

Similarly, by (63b) and the monotonicity of $\beta(\cdot)$ and $\beta'(\cdot)$,

$$\begin{aligned} i \in [\nu\mathrm{med}_+\{t\}] \Longrightarrow H(\tilde{F}_i|\tilde{G}_i) &< \beta'(1/2 + \zeta/2 - \xi) \\ &= 1 + \zeta - 2\xi - 1 + \xi \\ &< \zeta. \end{aligned}$$

We have proved the claim for $\tilde{\mathcal{H}} = h(\mathbf{c})$. The general case follows by monotonicity. That is, recall that the derivation above relies on repeated applications of the functions $\alpha, \alpha', \beta, \beta'$ in Corollary 18. These functions are monotone. Thus, since we have proved (104a) and (105a) for $i \in [\mathrm{med}_+\{t\}]$ when $\tilde{\mathcal{H}} = h(\mathbf{c})$, this must also be the case for $\tilde{\mathcal{H}} \le h(\mathbf{c})$. The case $\tilde{\mathcal{H}} \ge h(\mathbf{c})$ follows similarly. ■

The following corollary of Proposition 29 is the analog of Corollary 19 for the cascade case.

Corollary 30. *For a given $\zeta > 0$, let t, $\mathbf{c}$, L_0, M_0, and n_{th} be as in Proposition 29. Then, under the same setting as Proposition 29, for any $0 \le \tau \le 1$ and $n \ge n_{\mathrm{th}}$ we have*

$$\tilde{\mathcal{H}} \le h(\mathbf{c}) + \frac{\tau}{2^t} \Rightarrow H(\tilde{F}_i|\tilde{G}_i) < 2\zeta + \tau, \qquad \forall i \in [\mathrm{med}_+\{t\}], \tag{108a}$$

$$\tilde{\mathcal{H}} \ge h(\mathbf{c}) - \frac{\tau}{2^t} \Rightarrow H(\tilde{F}_i|\tilde{G}_i) > 1 - 2\zeta - \tau, \forall i \in [\mathrm{med}_-\{t\}]. \tag{108b}$$

Proof: Recall that in the cascade construction, when moving from stage $i < t$ to stage $i+1$, we operate only on the new medial indices in the active set σ_i, defined in (90). The remaining new medial indices belong to the set $\bar{\sigma}_i$, defined as:

$$\bar{\sigma}_i = \begin{cases} [\nu\mathrm{med}_-\{i\}], & \text{if } c_i = 1, \\ [\nu\mathrm{med}_+\{i\}], & \text{if } c_i = 0. \end{cases}$$

By assumption, we are in a BI-process setting. By Lemma 13, the conditional entropy corresponding to any index in σ_i is the same, and hence we denote it by a_i. Similarly, we denote by b_i the conditional entropy corresponding to an arbitrary index in $\bar{\sigma}_i$. For stage t, we denote by a_t^- and a_t^+ the conditional entropies corresponding to indices in $[\nu\mathrm{med}_-\{t\}]$ and $[\nu\mathrm{med}_+\{t\}]$, respectively. Furthermore, conservation of conditional entropy holds, by Corollary 14. That is,

$$\begin{aligned} a_1 + b_1 &= 2\tilde{\mathcal{H}}, \\ a_i + b_i &= 2a_{i-1}, \quad 2 \le i < t, \\ a_t^- + a_t^+ &= 2a_{t-1}. \end{aligned}$$

From the above, we easily get by induction that

$$2^t \tilde{\mathcal{H}} = (a_t^- + a_t^+) + \sum_{i=1}^{t-1} 2^{t-i} b_i.$$

From the above and (93), we have

$$\tilde{\mathcal{H}} - h(\mathbf{c}) = \frac{a_t^- + a_t^+ - 1 + \sum_{i=1}^{t-1} 2^{t-i}(b_i - c_i)}{2^t}.$$

We now prove (108a). If $\tilde{\mathcal{H}} \le h(\mathbf{c})$, then the result follows trivially from (104a) in Proposition 29. Hence, assume that

$$h(\mathbf{c}) < \tilde{\mathcal{H}} \le h(\mathbf{c}) + \tau/2^t. \tag{109}$$

By (109), (105b), and the definition of $\bar{\sigma}_i$, observe that if $c_i = 1$ then $b_i > 1 - \zeta/2^t$. Moreover, by (104b) and (109), $a_t^- > 1 - \zeta$. Thus,

$$\begin{aligned} \tau &\ge 2^t(\tilde{\mathcal{H}} - h(\mathbf{c})) \\ &= a_t^- + a_t^+ - 1 + \sum_{i=1}^{t-1} 2^{t-i}(b_i - c_i) \\ &= \left(a_t^+ + \sum_{i,c_i=0} 2^{t-i} b_i\right) + \left(a_t^- - 1 + \sum_{i,c_i=1} 2^{t-i}(b_i - 1)\right) \\ &> \left(a_t^+ + \sum_{i,c_i=0} 2^{t-i} b_i\right) - \left(1 + \sum_{i,c_i=1} 2^{-i}\right)\zeta \\ &> \left(a_t^+ + \sum_{i,c_i=0} 2^{t-i} b_i\right) - 2\zeta. \end{aligned}$$

Rearranging the above yields

$$a_t^+ + \sum_{i,c_i=0} 2^{t-i} b_i < 2\zeta + \tau.$$

By the non-negativity of conditional entropy, this implies that $a_t^+ < 2\zeta + \tau$ and for i such that $c_i = 0$, we have $b_i < 2\zeta + \tau$. Finally, recalling the cascade construction, the definition of $[\text{med}_+\{t\}]$, and the definition of $\bar{\sigma}_i$, the conditional entropy of any index in $[\text{med}_+\{t\}]$ is either a_t^+ or some b_i where i is such that $c_i = 0$. This yields (108a).

The proof for (108b) is similar. If $\tilde{\mathcal{H}} \geq h(\mathbf{c})$, then the result follows trivially from (104b) in Proposition 29. Hence, assume that

$$h(\mathbf{c}) - \tau/2^t \leq \tilde{\mathcal{H}} < h(\mathbf{c}). \tag{110}$$

By (110), (105a), and the definition of $\bar{\sigma}_i$, observe that if $c_i = 0$ then $b_i < \zeta/2^t$. Moreover, by (104a) and (110), $a_t^+ < \zeta$. Thus,

$$\begin{aligned}
-\tau &\leq 2^t(\tilde{\mathcal{H}} - h(\mathbf{c})) \\
&= a_t^- + a_t^+ - 1 + \sum_{i=1}^{t-1} 2^{t-i}(b_i - c_i) \\
&= \left(a_t^+ + \sum_{i,c_i=0} 2^{t-i} b_i\right) + \left(a_t^- - 1 + \sum_{i,c_i=1} 2^{t-i}(b_i - 1)\right) \\
&< \left(\zeta + \sum_{i,c_i=0} 2^{-i}\zeta\right) + \left(a_t^- - 1 + \sum_{i,c_i=1} 2^{t-i}(b_i - 1)\right) \\
&< 2\zeta + \left((a_t^- - 1) + \sum_{i,c_i=1} 2^{t-i}(b_i - 1)\right).
\end{aligned}$$

Rearranging the above yields

$$-2\zeta - \tau < (a_t^- - 1) + \sum_{i,c_i=1} 2^{t-i}(b_i - 1).$$

Since conditional entropy is upper-bounded by 1, we have $(a_t^- - 1) \leq 0$ and $(b_i - 1) \leq 0$. The above inequality thus implies that $(a_t^- - 1) > -2\zeta - \tau$ and for i such that $c_i = 1$, we have $(b_i - 1) > -2\zeta - \tau$. Finally, recalling the cascade construction, the definition of $[\text{med}_-\{t\}]$, and the definition of $\bar{\sigma}_i$, the conditional entropy of any index in $[\text{med}_-\{t\}]$ is either a_t^- or some b_i where i is such that $c_i = 1$. Simple rearranging yields (108b). ■

The following is the analog of Proposition 22 for the cascade case. Similar to Proposition 22, we state the following for a forgetful FAIM-derived s/o-process, $X_\star \rightarrowtail Y_\star$, that satisfies (66) for some sequences $\bar{\epsilon}_L$, $\bar{\psi}_M$, and $\bar{\phi}_M$.

Proposition 31. *Let the number of stages of the cascade t, and the binary vector $\mathbf{c}$ of length $t-1$ be given. Fix $n \geq 0$, $\varepsilon_1 > 0$, and $0 < \varepsilon_2 < \frac{1}{6}$. There exist L_{th} and M_{th} such that for any $L_0 \geq L_{\text{th}}, M_0 \geq M_{\text{th}}$, a $(t, \mathbf{c}; L_0, M_0, n)$-cascade satisfies:*

$$\begin{aligned}
|H(F_i|G_i) - H(\tilde{F}_i|\tilde{G}_i)| &\leq 2\varepsilon_1 + \frac{\varepsilon_2}{2} - 3\varepsilon_2 \log\frac{3\varepsilon_2}{2} \\
&< 2\varepsilon_1 + \sqrt{8\varepsilon_2}.
\end{aligned}$$

Furthermore, we have

$$|H(F_i|G_i, S_0, S_{N_n}) - H(\tilde{F}_i|\tilde{G}_i)| \leq 2\varepsilon_1 + \frac{\varepsilon_2}{2} - 3\varepsilon_2 \log\frac{3\varepsilon_2}{2}.$$

Proof: The proof follows along the same lines as Proposition 22. In the cascade case, the base-vector $\mathbf{b}$ of index i is of length $2^{n \cdot t}$ as opposed to 2^n in the non-cascade case. This vector holds a single medial index from each of the b-blocks on the RHS of the transform, just as for the non-cascade case. All that remains are minor adaptations to the proof, to account for this change. That is, throughout the proof, including in the underlying lemmas, we make the following changes.

- Replace 2^n and 2^{-n} with $2^{n \cdot t}$ and $2^{-n \cdot t}$, respectively.
- Replace N_n with $N^{\{t\}}$.
- Extend $\tilde{f}_{n,i}$ and $\tilde{g}_{n,i}$ to the cascade case according to the construction in Section V-D.

This completes the proof. ■

Proof of Theorem 28: Choose $\varepsilon_1 > 0$ and $0 < \varepsilon_2 < \frac{1}{6}$ small enough such that

$$\zeta \triangleq \eta - 2^{t+1}\varepsilon_1 - (2\varepsilon_1 + \sqrt{8\varepsilon_2}) > 0. \tag{111}$$

For example, one may take

$$\begin{aligned}
\varepsilon_1 &< \frac{\eta}{4(1+2^t)}, \\
\varepsilon_2 &< \frac{\eta^2}{32}.
\end{aligned}$$

Take n_{th} large enough so that Proposition 29 holds with ζ as above. Recall that Proposition 29 holds for any L_0 and M_0, so we are free to set them as desired.

By Proposition 31, for n_{th}, ε_1, and ε_2 above, there exist L_{th} and M_{th} such that (68) holds for $L_0 \geq L_{\text{th}}$ and $M_0 \geq M_{\text{th}}$. That is,

$$-(2\varepsilon_1 + \sqrt{8\varepsilon_2}) \leq H(F_i|G_i) - H(\tilde{F}_i|\tilde{G}_i) \leq (2\varepsilon_1 + \sqrt{8\varepsilon_2}). \tag{112}$$

In fact, we choose $L_0 \geq L_{\text{th}}$ as in the proof of Lemma 23 (adapted to the cascade case as detailed in the proof of Proposition 31). This ensures that the L_0-forgetfulness of the s/o-process is upper-bounded by ε_1. Thus, by Corollary 12, (43) holds with $\epsilon \leq \varepsilon_1$, so that

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

Hence, if $\mathcal{H}(X_\star|Y_\star) \leq h(\mathbf{c})$ then $\tilde{\mathcal{H}} \leq h(\mathbf{c}) + 2\varepsilon_1$ and if $\mathcal{H}(X_\star|Y_\star) \geq h(\mathbf{c})$ then $\tilde{\mathcal{H}} \geq h(\mathbf{c}) - 2\varepsilon_1$. Consequently, by Corollary 30 with $\tau = 2^{t+1}\varepsilon_1$, if $n \geq n_{\text{th}}$ then

$$\begin{aligned}
\mathcal{H}(X_\star|Y_\star) \leq h(\mathbf{c}) &\Rightarrow H(\tilde{F}_i|\tilde{G}_i) < 2\zeta + \tau, && \forall i \in [\text{med}_+(n)], \\
\mathcal{H}(X_\star|Y_\star) \geq h(\mathbf{c}) &\Rightarrow H(\tilde{F}_i|\tilde{G}_i) > 1 - 2\zeta - \tau, && \forall i \in [\text{med}_-(n)].
\end{aligned}$$

Combining the above with (111) and (112) we obtain that for $n \geq n_{\text{th}}$,

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

This completes the proof. ■

VI. Decoding the Universal Polar Code

The universal polar code consists of a concatenation of a BST cascade and Arıkan's seminal codes, that is fast transforms. Ultimately, the code consists 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 connected. Therefore, both the slow and fast polarization stages are decoded using successive-cancellation decoding, performed in lockstep.

Specifically, the decoder estimates the transformed bits (the $\hat{U}$ in Figure 10) 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 [24] 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 seminal 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 $\Lambda = N \cdot \hat{N}$ (see Section III-C), so its decoding complexity using successive-cancellation trellis decoding is $O(|\mathcal{S}|^3\Lambda \cdot \log(\Lambda))$. In the proof of Theorem 1 below, we show that the overall decoding error of this scheme is upper-bounded by $2^{-\Lambda^\beta}$ for any $\beta < 1/2$ and $\hat{N}$ large enough.

As we will see in Section VIII, the decoding performance may be enhanced by using a successive cancellation list decoder. Such a decoder, tailored to the universal polar coding scheme of this paper is described in [25]. For a list of size L, the decoding complexity increases by a factor of L, with respect to (plain) successive cancellation decoding.

To prove Theorem 1 we will need some notation for the inputs and outputs of the slow and fast stages of the overall transform. The notation is illustrated in Figure 10. The construction consists of a layer of $\hat{N}$ copies of a BST cascade, each of length N, which is concatenated to a layer of N fast transforms, each of length $\hat{N}$. A vector comprising the jth output of each BST cascade is the input to fast transform j. Let $1 \le \ell \le \hat{N}$, $1 \le i \le \hat{N}$, and $1 \le j \le N$. Denote

$$X_j^{(\ell)} = X_{j+(\ell-1)N},$$
$$Y_j^{(\ell)} = Y_{j+(\ell-1)N},$$

and their vector versions

$$\boldsymbol{X}_\ell = \begin{bmatrix} X_1^{(\ell)} & X_2^{(\ell)} & \cdots & X_N^{(\ell)} \end{bmatrix},$$
$$\boldsymbol{Y}_\ell = \begin{bmatrix} Y_1^{(\ell)} & Y_2^{(\ell)} & \cdots & Y_N^{(\ell)} \end{bmatrix}.$$

That is, the s/o-process relevant for BST cascade ℓ is $\boldsymbol{X}_\ell \rightarrowtail \boldsymbol{Y}_\ell$. Output j of BST cascade ℓ is $F_j^{(\ell)}$. The input to fast transform j is the vector

$$\boldsymbol{F}_j = \begin{bmatrix} F_j^{(1)} & F_j^{(2)} & \cdots & F_j^{(\hat{N})} \end{bmatrix}.$$

Output i of fast transform j is $\hat{U}_i^{(j)} = \hat{U}_{i+(j-1)\hat{N}}$. The overall output of fast transform j is the vector

$$\hat{\boldsymbol{U}}_j = \begin{bmatrix} \hat{U}_1^{(j)} & \hat{U}_2^{(j)} & \cdots & \hat{U}_{\hat{N}}^{(j)} \end{bmatrix}.$$

Our notation for a vector of ordered elements (see Section II) carriers over to an ordered vector of vectors, e.g., $\hat{\boldsymbol{U}}_a^b = \begin{bmatrix} \hat{\boldsymbol{U}}_a & \hat{\boldsymbol{U}}_{a+1} & \cdots & \hat{\boldsymbol{U}}_b \end{bmatrix}$. Moreover, $\hat{U}_1^{N\hat{N}} = \hat{\boldsymbol{U}}_1^N$.

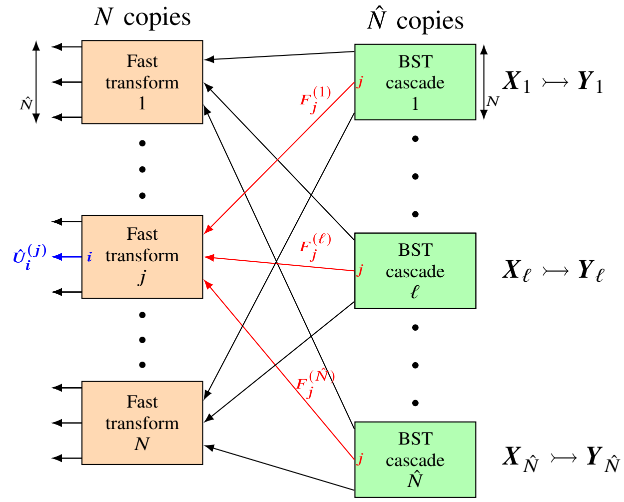

Fig. 10. Notation for inputs and outputs of the slow and fast stages in our universal construction.

Note that under this notation, the decoder decodes the bit vectors $\hat{\boldsymbol{U}}_j$, $j = 1, \ldots, N$ in order using successive-cancellation trellis decoding. Specifically, it first decodes $\hat{\boldsymbol{U}}_1$, then $\hat{\boldsymbol{U}}_2$, and so on. The decoding order in a vector $\hat{\boldsymbol{U}}_j$ is as expected: $\hat{U}_1^{(j)}$, then $\hat{U}_2^{(j)}$, and so forth up to $\hat{U}_{\hat{N}}^{(j)}$. Namely, when decoding $\hat{U}_i^{(j)}$, we do this after having decoded $\hat{\boldsymbol{U}}_1^{j-1}$ as well as $\hat{U}_1^{(j)}, \hat{U}_2^{(j)}, \ldots, \hat{U}_{i-1}^{(j)}$.

Proof of Theorem 1: Our proof is divided into four parts:

I Defining the coding scheme, i.e., the indices on which information bits are transmitted.
II Proving item 3 in the theorem statement: encoding and decoding complexity.
III Proving item 2 in the theorem statement: vanishing error probability and the error exponent.
IV Proving item 1 in the theorem statement: the code rate R is achievable.

Part I – Coding Scheme Definition

Our sequence of codes is parametrized by $R < I^*$ and $\beta < 1/2$, both given in the theorem statement. It is based on the universal construction of this paper. That is, it consists of a concatenation of a cascade of BSTs and a fast transform. To fully specify a member of this sequence, we must define the following:

- The slow transform parameters: $t, \mathbf{c}, L_0, M_0$, and n.
- The number of stages in the fast transform: $\hat{n}$.
- The set of indices A over which information bits are transmitted.

The sequence is formed by increasing $\hat{n}$ and keeping the slow transform parameters fixed. The parameter $\hat{n}$ must be large enough; conditions on its size are given in this part, as well as in Parts III and IV. All these conditions must be met.

Let $\mathcal{C}$ be the family of s/o-processes. If $I^* = 0$, the theorem is trivially correct; hence, we assume that $I^* > 0$. Let $\bar{\epsilon}_L$, $\bar{\psi}_M$, and $\bar{\phi}_M$ be sequences such that (66) holds for any s/o-process in $\mathcal{C}$. Every s/o-process $X_\star \rightarrowtail Y_\star$ in $\mathcal{C}$ has the same input distribution; we denote the entropy rate of the input process by $\mathcal{H}(X_\star)$. Denote

$$h^* = \mathcal{H}(X_\star) - I^*. \tag{113}$$

By definition of I^*, any s/o-process $X_\star \rightarrowtail Y_\star$ in $\mathcal{C}$ satisfies $I^* \leq \mathcal{H}(X_\star) - \mathcal{H}(X_\star|Y_\star)$. In other words, we always have

$$\mathcal{H}(X_\star|Y_\star) \leq h^*. \tag{114}$$

Denote

$$\delta = I^* - R. \tag{115}$$

Looking forward, in part III we will use the union bound to upper-bound the probability of error. Thus, we will require fast polarization, for which we turn to Proposition 49. All processes in $\mathcal{C}$ satisfy the Bhattacharyya recursion (166) with

$$\kappa = 2\bar{\psi}_0, \tag{116}$$

see (168) and (66b). Next, fix β' such that

$$\beta < \beta' < 1/2 \tag{117}$$

and let

$$\delta' = \delta/5. \tag{118}$$

By Proposition 49, there exist η and n_0 such that if $Z_0 \leq \eta$, then (170) below holds with β' and δ' in place of β and δ. This will be utilized in Part IV. Parameters η and n_0 are used as follows.

- Parameter η will be the monopolarization goal for the BST cascade in Theorem 28.
- Parameter n_0 will be one of the lower bounds on the number of fast polarization steps, $\hat{n}$.

To apply Theorem 28, we now set the parameters t and $\mathbf{c}$. We set t and $\mathbf{c}$ such that

$$h^* \leq h(\mathbf{c}) \leq h^* + \delta'. \tag{119}$$

Indeed, this can be done by (93), recalling that $I^* > 0$ implies by (113) that $h^* < 1$.

Recall from Definition 12 that a cascade of BSTs is defined by five parameters: $t, \mathbf{c}, L_0, M_0$, and n. We have already set t and $\mathbf{c}$. For the remaining parameters, we will utilize Theorem 28. Namely, given $t, \mathbf{c}$, and η, there exist $L_{\text{th}}, M_{\text{th}}$, and n_{th} (see Remark 7) such that if $L_0 \geq L_{\text{th}}$, $M_0 \geq M_{\text{th}}$, and $n \geq n_{\text{th}}$ then a $(t, \mathbf{c}; L_0, M_0, n)$-cascade is $(\eta, [\text{med}_+\{t\}], [\text{med}_-\{t\}])$-monopolarizing, for any s/o-process in $\mathcal{C}$. Specifically, since $\mathcal{H}(X_\star|Y_\star) \leq h(\mathbf{c})$ by (114) and (119), the medial-plus indices are the ones that monopolarize. Thus, due to Proposition 49, after the fast transform almost all of them will have a very low Bhattacharyya parameter.

We set $n = n_{\text{th}}$. To set L_0 and M_0, we further call upon Lemma 26. Namely, we set $L_0 = L_{\text{th}}$ and take $M_0 \geq M_{\text{th}}$ large enough such that the fraction of medial-plus indices out of all indices in our BST cascade is at least

$$\frac{|[\text{med}_+\{t\}]|}{N} \geq 1 - h(\mathbf{c}) - \delta'. \tag{120}$$

We can do this due to the combination of (94) and (95), and by recalling the observation following (95).

Up to this point, we have defined all the parameters of the BST cascade: $t, \mathbf{c}, L_0, M_0, n$. Recall that the sequence is formed by increasing $\hat{n}$, and that $\hat{n}$ must satisfy a set of conditions that we have yet to fully specify. For now, consider some $\hat{n}$ large enough.

Using (92), let $N = (2L_0 + M_0)M_0^{t-1}2^{nt}$ and denote

$$\hat{N} = 2^{\hat{n}}. \tag{121}$$

The total codelength is thus

$$\Lambda = N \cdot \hat{N} = (2L_0 + M_0)M_0^{t-1}2^{nt} \cdot 2^{\hat{n}}. \tag{122}$$

We now define the set D as all indices $k = i + (j-1)\hat{N}$, $1 \leq i \leq \hat{N}$, $1 \leq j \leq N$ such that the following conditions hold:

1) $j \in [\text{med}_+\{t\}]$, where we recall that $[\text{med}_+\{t\}]$ is the medial-plus set for our $(t, \mathbf{c}; L_0, M_0, n)$-cascade of BSTs;
2) $\bar{Z}_{\hat{n}} \leq 2^{-\hat{N}^{\beta'}}$, where $\bar{Z}_0 = \eta$, and the process $\bar{Z}$ satisfies (166) with equality instead of inequality, $\kappa = 2\bar{\psi}_0$ (see (116)), and we take B_{m+1} as the mth bit in the binary representation of i.

Remark 9. Observe that if $k \in \mathsf{D}$, then for any s/o-process in $\mathcal{C}$, we have $Z(\hat{U}_k|\hat{U}_1^{k-1}, Y_1^\Lambda) \leq 2^{-\hat{N}^{\beta'}}$. This follows from the second condition in the definition of D, which upper-bounds the Bhattacharyya process for any s/o-process in $\mathcal{C}$.

We now partition D into three disjoint sets, A, B, and C.

$$\mathsf{A} = \left\{k \in \mathsf{D} : K(\hat{U}_k|\hat{U}_1^{k-1}) \leq 2^{-\hat{N}^{\beta'}}\right\},$$
$$\mathsf{B} = \left\{k \in \mathsf{D} : K(\hat{U}_k|\hat{U}_1^{k-1}) \geq 1 - 2^{-\hat{N}^{\beta'}}\right\},$$
$$\mathsf{C} = \left\{k \in \mathsf{D} : 2^{-\hat{N}^{\beta'}} < K(\hat{U}_k|\hat{U}_1^{k-1}) < 1 - 2^{-\hat{N}^{\beta'}}\right\},$$

where K is the total variation distance, see [26, Definition 3]. Thus,

$$|\mathsf{D}| = |\mathsf{A}| + |\mathsf{B}| + |\mathsf{C}|. \tag{123}$$

Following the Honda-Yamamoto scheme [24], we take A as the set of information indices.

Part II – Complexity

We have already discussed the decoding complexity of our scheme in the beginning of this section, and showed that it is $O(|\mathcal{S}|^3\Lambda \cdot \log(\Lambda))$. This is also the encoding complexity, since encoding uses the successive cancellation trellis algorithm [16] as well. This proves item 3 in the theorem statement.

Part III – Error Probability

To upper-bound the probability of error, we will use the union bound. For this, we will need a second condition on $\hat{n}$: it is large enough such that

$$2^{-\hat{N}^{\beta'}} \leq \frac{1}{2\Lambda}2^{-\Lambda^\beta}. \tag{124}$$

It is possible to satisfy (124) for all $\hat{n}$ large enough since $\beta' > \beta$ by (117), N is fixed, and $\Lambda = N \cdot \hat{N}$ by (122).

Recall that the input distribution is fixed over the set $\mathcal{C}$. Denote this input distribution by $P_{X_1^\Lambda}$. Since $\hat{U}_1^\Lambda$ is the result of a transform we have specified over X_1^Λ, we denote the corresponding distribution of $\hat{U}_1^\Lambda$ as $P_{\hat{U}_1^\Lambda}$.

The claim on the probability of error follows from [24]; for completeness we show this directly here.

- We encode $\hat{u}_k$ sequentially.

– If $k \in \mathsf{A}$, encode the next information bit into $\hat{u}_k$. We assume that the information bits are i.i.d. Bernoulli$(1/2)$.
– If $k \notin \mathsf{A}$, we draw $\hat{u}_k$ according to a binary random variable with distribution $P_{\hat{U}_k|\hat{U}_1^{k-1}}(u|\hat{u}_1^{k-1})$.

- This results in a $\hat{u}_1^\Lambda$ that is sampled from a distribution $Q_{\hat{U}_1^\Lambda}$ that is different from the distribution $P_{\hat{U}_1^\Lambda}$.
- By Remark 9 and (124), had $\hat{u}_1^\Lambda$ been sampled from $P_{\hat{U}_1^\Lambda}$, the union bound would yield a total probability of error of at most $\frac{1}{2}2^{-\Lambda^\beta}$.
- We abuse notation and denote by $P(\hat{u}_1^\Lambda, y_1^\Lambda)$ the joint probability of $\hat{u}_1^\Lambda$ sampled from $P_{\hat{U}_1^\Lambda}$ and y_1^Λ the resulting channel output. Similarly, $Q(\hat{u}_k, y_1^\Lambda)$ denotes this joint probability if $\hat{u}_1^\Lambda$ were sampled from $Q_{\hat{U}_1^\Lambda}$. Under this notation, from the previous bullet, we have

$$\sum_{\hat{u}_1^\Lambda, y_1^\Lambda} P(\hat{u}_1^\Lambda, y_1^\Lambda) \cdot \left[\mathrm{Dec}(y_1^\Lambda) \neq \hat{u}_1^\Lambda\right] \leq \frac{1}{2}2^{-\Lambda^\beta}, \tag{125}$$

where $[\cdot]$ is the Iverson notation [27, p. 11]: equal to 1 if the condition holds and to 0 otherwise, and 'Dec' is the successive cancellation decoder.
- The true probability of error of this scheme, i.e., under Q, is given by

$$\begin{aligned}
P_{\mathrm{e}} &= \sum_{\hat{u}_1^\Lambda, y_1^\Lambda} Q(\hat{u}_1^\Lambda, y_1^\Lambda) \cdot \left[\mathrm{Dec}(y_1^\Lambda) \neq \hat{u}_1^\Lambda\right] \\
&= \sum_{\hat{u}_1^\Lambda, y_1^\Lambda} P(\hat{u}_1^\Lambda, y_1^\Lambda) \cdot \left[\mathrm{Dec}(y_1^\Lambda) \neq \hat{u}_1^\Lambda\right] \\
&\quad + \sum_{\hat{u}_1^\Lambda, y_1^\Lambda} \left(Q(\hat{u}_1^\Lambda, y_1^\Lambda) - P(\hat{u}_1^\Lambda, y_1^\Lambda)\right) \cdot \left[\mathrm{Dec}(y_1^\Lambda) \neq \hat{u}_1^\Lambda\right] \\
&\leq \frac{1}{2}2^{-\Lambda^\beta} + \sum_{\hat{u}_1^\Lambda, y_1^\Lambda} \left|Q(\hat{u}_1^\Lambda, y_1^\Lambda) - P(\hat{u}_1^\Lambda, y_1^\Lambda)\right|,
\end{aligned}$$

where the inequality is by (125), and since $[\cdot] \leq 1$.
- Following [24, eq. 57] (see also [28, Lemma 3.5]), we can bound the second term above as

$$\begin{aligned}
&\sum_{\hat{u}_1^\Lambda, y_1^\Lambda} \left|Q(\hat{u}_1^\Lambda, y_1^\Lambda) - P(\hat{u}_1^\Lambda, y_1^\Lambda)\right| \\
&\stackrel{(a)}{\leq} \sum_{k \in \mathsf{A}} \sum_{\hat{u}_1^k} P(\hat{u}_1^{k-1}) \left|Q(\hat{u}_k|\hat{u}_1^{k-1}) - P(\hat{u}_k|\hat{u}_1^{k-1})\right| \\
&\stackrel{(b)}{=} \sum_{k \in \mathsf{A}} \sum_{\hat{u}_1^{k-1}} P(\hat{u}_1^{k-1}) \sum_{\hat{u}_k} \left|\frac{1}{2} - P(\hat{u}_k|\hat{u}_1^{k-1})\right| \\
&\stackrel{(c)}{=} \sum_{k \in \mathsf{A}} \sum_{\hat{u}_1^{k-1}} P(\hat{u}_1^{k-1}) \left|P(0|\hat{u}_1^{k-1}) - P(1|\hat{u}_1^{k-1})\right| \\
&\stackrel{(d)}{=} \sum_{k \in \mathsf{A}} K(\hat{U}_k|\hat{U}_1^{k-1}) \\
&\stackrel{(e)}{\leq} |\mathsf{A}| 2^{-\hat{N}^{\beta'}} \\
&\leq \Lambda \cdot 2^{-\hat{N}^{\beta'}} \\
&\stackrel{(f)}{\leq} \frac{1}{2}2^{-\Lambda^\beta},
\end{aligned}$$

where in (a) we abused notation and used P and Q to denote $P_{\hat{U}_1^{k-1}}$, $Q_{\hat{U}_k|\hat{U}_1^{k-1}}$, and $P_{\hat{U}_k|\hat{U}_1^{k-1}}$; in (b) we recalled from the topmost bullet that for $k \in \mathsf{A}$ we draw $\hat{u}_k$ from a Bernoulli$(1/2)$ random variable; in (c) we used the equality $1/2 = (P(0|\hat{u}_1^{k-1}) + P(1|\hat{u}_1^{k-1}))/2$; in (d) we used [14, Definition 3]; (e) follows from the definition of A; finally, (f) is by (124).
- Combining the above two bullets, we obtain

$$P_{\mathrm{e}} \leq 2^{-\Lambda^\beta}.$$

This proves item 2 in the theorem statement.

Part IV – Rate

Our goal is to show that

$$|\mathsf{A}| \geq \Lambda R. \tag{126}$$

The size of D is lower-bounded by

$$\begin{aligned}
|\mathsf{D}| &\stackrel{(a)}{\geq} N \cdot (1 - h(\mathbf{c}) - \delta') \cdot \hat{N} \cdot (1 - \delta') \\
&\stackrel{(b)}{=} \Lambda(1 - h(\mathbf{c}) - \delta') \cdot (1 - \delta') \\
&\stackrel{(c)}{>} \Lambda(1 - h(\mathbf{c}) - 2\delta') \\
&\stackrel{(d)}{\geq} \Lambda(1 - h^* - 3\delta') \\
&\stackrel{(e)}{=} \Lambda(I^* + 1 - \mathcal{H}(X_\star) - 3\delta'),
\end{aligned} \tag{127}$$

where

- (a) is by (120) and by (170) with β' and δ' in place of β and δ, recalling (116) and the discussion following (118).
- (b) is by (122).
- (c) is since $\delta' > 0$. See (118), (115), and the text preceding (115).
- (d) is by (119).
- (e) is by (113).

To lower-bound the size of A, we define the sets A', B', C', which are defined similarly to A, B, and C, but we do not limit the indices k to be from D. That is,

$$\begin{aligned}
\mathsf{A}' &= \left\{1 \leq k \leq \Lambda : K(\hat{U}_k|\hat{U}_1^{k-1}) \leq 2^{-\hat{N}^{\beta'}}\right\}, \\
\mathsf{B}' &= \left\{1 \leq k \leq \Lambda : K(\hat{U}_k|\hat{U}_1^{k-1}) \geq 1 - 2^{-\hat{N}^{\beta'}}\right\}, \\
\mathsf{C}' &= \left\{1 \leq k \leq \Lambda : 2^{-\hat{N}^{\beta'}} < K(\hat{U}_k|\hat{U}_1^{k-1}) < 1 - 2^{-\hat{N}^{\beta'}}\right\}.
\end{aligned}$$

Clearly, $|\mathsf{B}| \leq |\mathsf{B}'|$ and $|\mathsf{C}| \leq |\mathsf{C}'|$, thus

$$|\mathsf{B}| + |\mathsf{C}| \leq |\mathsf{B}'| + |\mathsf{C}'|. \tag{128}$$

We now prove that

$$|\mathsf{B}'| + |\mathsf{C}'| \leq \Lambda(1 - \mathcal{H}(X_\star) + 2\delta'). \tag{129}$$

First, let us show that for $\hat{n}$ large enough,

$$|\mathsf{C}'| \leq \Lambda\delta'. \tag{130}$$

Recall from Figure 10 that the transform consists of a concatenation of copies of BST cascades with fast transforms. Specifically, for fixed $1 \leq j \leq N$, all indices of the form $k = i + (j-1)\hat{N}$, $1 \leq i \leq \hat{N}$ belong to the same fast transform. We claim that for a fixed j there exists an $\hat{n}$ large enough such that the fraction of such indices k belonging to C' is at most δ'. Indeed, this follows from fast polarization results in [26]. To see this, first denote $Z_k = Z(\hat{U}_k|\hat{U}_1^{k-1})$ and $K_k = K(\hat{U}_k|\hat{U}_1^{k-1})$. From [26, (4a) and (4b)], if $Z_k \le 2^{-\hat{N}^\beta}$ then

$$K_k \ge 1 - Z_k \ge 1 - 2^{-\hat{N}^\beta}.$$

Combining the above with (28) in [26, Theorem 7] and the first two displayed equations in the proof of [26, Theorem 13], we see the claim.

Now, to see (130), note that since we have fixed N, the number of possible indices j is finite. Hence, we take $\hat{n}$ as being at least the maximum $\hat{n}$ over all such j. This yields (130).

We now show that for $\hat{n}$ large enough,

$$|\mathsf{A}'| \ge \Lambda\left(\mathcal{H}(X_\star) - 2\delta'\right). \tag{131}$$

Indeed, this follows by

$$\begin{aligned}
&\Lambda\mathcal{H}(X_\star) \\
&\overset{\text{(a)}}{\le} \sum_{k=1}^{\Lambda} H(\hat{X}_k|\hat{X}_1^{k-1}) \\
&\overset{\text{(b)}}{=} \sum_{k=1}^{\Lambda} H(\hat{U}_k|\hat{U}_1^{k-1}) \\
&= \sum_{k\in\mathsf{A}'} H(\hat{U}_k|\hat{U}_1^{k-1}) + \sum_{k\in\mathsf{B}'} H(\hat{U}_k|\hat{U}_1^{k-1}) + \sum_{k\in\mathsf{C}'} H(\hat{U}_k|\hat{U}_1^{k-1}) \\
&\overset{\text{(c)}}{\le} |\mathsf{A}'| + \sum_{k\in\mathsf{B}'} H(\hat{U}_k|\hat{U}_1^{k-1}) + |\mathsf{C}'| \\
&\overset{\text{(d)}}{\le} |\mathsf{A}'| + \sum_{k\in\mathsf{B}'} H(\hat{U}_k|\hat{U}_1^{k-1}) + \Lambda\delta' \\
&\overset{\text{(e)}}{\le} |\mathsf{A}'| + |\mathsf{B}'|\delta' + \Lambda\delta' \\
&\le |\mathsf{A}'| + \Lambda\delta' + \Lambda\delta' \\
&= |\mathsf{A}'| + 2\Lambda\delta',
\end{aligned}$$

where:

- (a) follows from [20, Theorem 4.2.2], which states that the summands in the series are a non-increasing sequence with limit $\mathcal{H}(X_\star)$.
- (b) follows since our overall transform is invertible.
- (c) follows since the summed entropies are upper bounded by 1.
- (d) follows from (130).
- To see (e), we utilize [26, (4c)]. That is, the inequality

  $$H(\hat{U}_k|\hat{U}_1^{k-1}) \le \sqrt{1 - (K(\hat{U}_k|\hat{U}_1^{k-1}))^2}.$$

  Since for every $k \in \mathsf{B}'$ we have $K(\hat{U}_k|\hat{U}_1^{k-1}) \ge 1 - 2^{-\hat{N}^{\beta'}}$, we can take $\hat{N}$ large enough such that $H(\hat{U}_k|\hat{U}_1^{k-1}) \le \delta'$.

Rearranging yields (131).

Finally, to obtain (129), observe that

$$\Lambda = |\mathsf{A}'| + |\mathsf{B}'| + |\mathsf{C}'|.$$

Thus, using (131) we obtain

$$|\mathsf{B}'| + |\mathsf{C}'| \le \Lambda(1 - \mathcal{H}(X_\star) + 2\delta'),$$

which is (129).

Having proved (129), we utilize (123) to conclude that

$$\begin{aligned}
|\mathsf{A}| &= |\mathsf{D}| - |\mathsf{B}| - |\mathsf{C}| \\
&\overset{\text{(a)}}{\ge} |\mathsf{D}| - |\mathsf{B}'| - |\mathsf{C}'| \\
&\overset{\text{(b)}}{\ge} |\mathsf{D}| - \Lambda(1 - \mathcal{H}(X_\star) + 2\delta') \\
&\overset{\text{(c)}}{\ge} \Lambda(I^* + 1 - \mathcal{H}(X_\star) - 3\delta') - \Lambda(1 - \mathcal{H}(X_\star) + 2\delta') \\
&= \Lambda(I^* - 5\delta') \\
&\overset{\text{(d)}}{=} \Lambda(I^* - \delta) \\
&\overset{\text{(e)}}{=} \Lambda R,
\end{aligned}$$

where:

- (a) is by (128).
- (b) is by (129).
- (c) is by (127).
- (d) is by (118).
- (e) is by (115).

Thus, the size of A is lower-bounded by R, proving item 1 in the theorem statement, and completing the proof. ■

VII. How to Construct Universal Polar Codes

In this section, we explain how one would construct universal polar codes in practice. Until this point, we have shown that a set of processes with memory that satisfy certain conditions can be decoded using a universal polar code. The code's parameters depend on these conditions and must be sufficiently large to attain a specified coding rate and error probability. A naive approach would be to set these parameters using Remark 7 and Section III-C. However, this might needlessly result in an impracticably large codelength — recall Example 5 that suggests a BST comprising of 40162 layers, whereas in fact just 10 layers suffice.

In what follows, for simplicity, we consider a universal transform whose slow stage is a single BST layer (i.e., a cascade with $t = 1$ layers). Generalizing to cascades with larger t is straightforward.

In a practical application, the code is to be used over some set of s/o-processes $\mathcal{C}$. All processes in $\mathcal{C}$ share the same fixed input distribution, which is known. We are further given a size constraint on the codelength.

To continue, we first decide the code parameters:

- The parameters L_0 and M_0 that respectively define the number of lateral and medial indices in the first level of the BST, see Section III-B.
- The number of levels of the BST, n. Thus, using L_0 and M_0 above, the length of each BST in the slow stage is $N = (2L_0 + M_0) \cdot 2^n$.
- The number of levels of the fast transform, $\hat{n}$. Recall that this also determines the number of copies of the BST in the slow stage, $\hat{N} = 2^{\hat{n}}$.

With these parameters set, the codelength is given by $\Lambda = N \cdot \hat{N}$, see (122). We can now construct the transform, but we must still determine the frozen and non-frozen indices, and thus set the code rate. To this end, we must supply some minimal information on $\mathcal{C}$. This information consists of a set of *bounds* on the following parameters, which must hold for *all processes* in $\mathcal{C}$.

- Upper bound on $H(X_0|Y_{-L_0}^{L_0}, X_{-L_0}^{-1})$.
- Recalling (35), an upper bound $\bar{\epsilon}_{L_0}$ on both $I(S_1; S_{L_0}|X_1^{L_0}, Y_1^{L_0})$ and $I(S_1; S_{L_0}|Y_1^{L_0})$.
- Upper bound $\bar{\psi}_{M_0-2}$ and lower bound $\bar{\phi}_{M_0-2}$, on ψ_{M_0-2} and ϕ_{M_0-2}, respectively, see Lemma 7.
- Upper bound $\bar{\psi}_0$ on ψ_0, which by (116) leads to an upper bound on κ.

In step 1b of the construction below, it will become apparent that we would like $\bar{\epsilon}_{L_0}$ to be close to 0 and both $\bar{\psi}_{M_0-2}$ and $\bar{\phi}_{M_0-2}$ to be close to 1. This can be achieved by taking L_0 and M_0 large enough. Next, we also must ensure that the fraction of medial indices, α_n in (18), is close to 1. Recalling Lemma 3, this can be achieved by ensuring that M_0 is sufficiently larger than L_0.

The above set of bounds is the key to computing upper bounds on the Bhattacharyya parameters of the s/o-pairs at the last level of the overall universal transform. These are part of what is required for determining the frozen set.

We must also take into consideration the fixed input process. We employ the Honda-Yamamoto [24] scheme to this end. Namely, we need to compute upper bounds on the total variation of the s/o-pairs at the last level of the overall universal transform for the input process. Practically, since the input process is fixed and independent of the channel, Monte-Carlo simulation may be used to obtain these bounds. These bounds, together with the above bounds on Bhattacharyya parameters, are used to determine the frozen set.

We now proceed via the following steps, which use the notation of Figure 10. Further recall that $G_j = (F_1^{j-1}, Y_1^N)$, see Definition 4.

1) Compute upper bounds on the conditional entropies of the s/o-pairs at the last level of each BST of the slow stage. That is, for each $1 \le j \le N$, compute upper bounds on $H(F_j|G_j) = H(F_j|F_1^{j-1}, Y_1^N)$. Note that we have dropped the index ℓ, since all BSTs are copies of one another, and we have assumed stationarity. This step consists of two substeps.
   a) For the first substep, we momentarily assume a block-independent process regime (recall Section V-B). Using Lemma 20 with the upper bound on $H(X_0|Y_{-L_0}^{L_0}, X_{-L_0}^{-1})$ in place of $\tilde{\mathcal{H}}$, calculate upper bounds on $H(\tilde{F}_j|\tilde{G}_j)$ (see (67c)) for each medial index j. Recall from Lemma 13 that at each level m of the BST, the entropy of all indices in $[\mathrm{med}_-(m)]$ is identical and similarly the entropy of all indices $[\mathrm{med}_+(m)]$ is also identical. The entropies of lateral indices are "stuck" and do not evolve further; at each level m of the transform two new lateral indices are generated, one from $[\mathrm{med}_-(m)]$ and one from $[\mathrm{med}_+(m)]$, so the bounds on their entropies stop evolving and are simply taken from the previous level.
   b) To remove the assumption from the previous substep, we use Proposition 22. That is, to compute upper bounds on $H(F_j|G_j)$, we use (68a). To this end, we compute an upper bound on ε_3, the right-hand side of (68a), using the bounds on the parameters above. Namely, in (71) we set $\varepsilon_1 \leftarrow \bar{\epsilon}_{L_0} \cdot 2^n$ and $\varepsilon_2 \leftarrow \max\{(\bar{\psi}_{M_0-2})^{2^n} - 1, 1 - (\bar{\phi}_{M_0-2})^{2^n}\}$, in line with (77) and (81), respectively. Observe that ε_3 bounds the difference between the results of this and the previous step; this difference may be made small by suitably increasing L_0 and M_0.
2) Compute upper bounds on the conditional entropies of the s/o-pairs at the last level of each fast transform of the slow stage, $Z\left(\hat{U}_i^{(j)} \,\middle|\, (\hat{U}_{i'}^{(j)})_{i'=1}^{i-1}, \hat{\boldsymbol{U}}_1^{j-1}, \boldsymbol{Y}_1^{\hat{N}}\right)$, for each $1 \le j \le N$ and $1 \le i \le \hat{N}$. We do this via the following substeps.
   a) Using [14, Lemma 1] on the result of step 1b, derive upper bounds on the Bhattacharyya parameters of the s/o-pairs entering the fast stage. That is, for each j, we compute the upper bound
   $$Z(F_j|F_1^{j-1}, Y_1^N) \le \sqrt{H(F_j|F_1^{j-1}, Y_1^N)}.$$
   b) From these, use (166) from Appendix A and the upper bound on κ to derive upper bounds on the Bhattacharyya parameters of the s/o-pairs at the last level of the fast stage. That is, for each $1 \le i \le \hat{N}$ and $1 \le j \le N$, these are upper bounds on
   $$Z\left(\hat{U}_i^{(j)} \,\middle|\, (\hat{U}_{i'}^{(j)})_{i'=1}^{i-1}, \hat{\boldsymbol{U}}_1^{j-1}, \boldsymbol{Y}_1^{\hat{N}}\right).$$

For the next step, we use the upper bounds on both the Bhattacharyya parameters and the total variations. That is, upper bounds on:
$$Z\left(\hat{U}_i^{(j)} \,\middle|\, (\hat{U}_{i'}^{(j)})_{i'=1}^{i-1}, \hat{\boldsymbol{U}}_1^{j-1}, \boldsymbol{Y}_1^{\hat{N}}\right),$$
$$K\left(\hat{U}_i^{(j)} \,\middle|\, (\hat{U}_{i'}^{(j)})_{i'=1}^{i-1}, \hat{\boldsymbol{U}}_1^{j-1}\right).$$

3) Take the non-frozen indices as those for which the sum of these upper bounds is small enough. That is, the word error rate is upper-bounded by
$$\sum_{i,j} \left(Z\left(\hat{U}_i^{(j)} \,\middle|\, (\hat{U}_{i'}^{(j)})_{i'=1}^{i-1}, \hat{\boldsymbol{U}}_1^{j-1}, \boldsymbol{Y}_1^{\hat{N}}\right) + K\left(\hat{U}_i^{(j)} \,\middle|\, (\hat{U}_{i'}^{(j)})_{i'=1}^{i-1}, \hat{\boldsymbol{U}}_1^{j-1}\right) \right),$$
where the sum is over the indices i, j for which $\hat{U}_i^{(j)}$ is non-frozen (contains an information bit).
Note that when the input distribution is uniform the total variation is always 0, thus one may consider only the sum of upper bounds on the Bhattacharyya parameter to determine the non-frozen set. Accordingly, in this case, the word error rate is upper-bounded by
$$\sum_{i,j} Z\left(\hat{U}_i^{(j)} \,\middle|\, (\hat{U}_{i'}^{(j)})_{i'=1}^{i-1}, \hat{\boldsymbol{U}}_1^{j-1}, \boldsymbol{Y}_1^{\hat{N}}\right),$$
where, again, the summation is over indices i, j for which $\hat{U}_i^{(j)}$ is non-frozen.

Remark 10. In a process with memory, the parameter ψ_0 is typically greater than 1 (see the discussion following Lemma 7), and thus κ from (116) is greater than 2. Thus, the upper bounds on the Bhattacharyya parameter may be non-informative for many s/o-pairs, leading to very low rate codes. To see this, consider the Bhattacharyya recursion (166) used in step 2b: in half of the steps, the recursion leads to the bound $Z_{n+1} \leq \kappa Z_n$. When κ is large, the right-hand side may quickly exceed 1, making this bound on the Bhattacharyya parameter non-informative. In practice, we can do much better. In the fast stage of the universal transform, we combine transformed s/o-pairs from different BSTs. These BSTs are "far apart:" at the beginning and end of each BST are enough lateral symbols that are essentially a buffer that ensures forgetfulness "takes place." Hence, practically we may assume that the two transformed s/o-pairs from different BSTs are independent and set $\psi_0 = 1$ (or $\kappa = 2$) in step 2b. This works well in practice, and we have followed this strategy for our numerical results below. The above heuristic can be made exact by adding an additional small buffer of symbols between BSTs. That is, due to the mixing property of FAIM processes, this buffer makes the BSTs as close to independent as desired. By Lemma 7, the mixing coefficients tend to 1 exponentially fast, so this buffer is negligible compared to the BST blocklength, incurring a vanishing rate loss.

We end this section by contrasting the construction of universal polar codes described above to that of non-universal polar codes. Here, by definition, we must construct the code to work for a *family* of s/o-processes. Thus, we cannot follow the process-specific technique described in [29].[7] Hence, we resort to using the above bounding techniques. It is interesting to note that for the universal setting, asymptotically (for large enough blocklengths), the use of these bounds does not incur a rate penalty. This is in contrast to non-universal polar codes.

VIII. Numerical Results

In this section, we provide simulation results for our universal polar code. These are given in Figures 11 and 12. The universal polar code was designed using the method in Section VII. Namely, we selected the following code parameters:

- $L_0 = 6$ and $M_0 = 40$,
- $n = 5$,
- $\hat{n} = 7$.

Thus, the code length is

$$N \cdot \hat{N} = (2L_0 + M_0) \cdot 2^{n+\hat{n}} = 212992.$$

We chose the memoryless uniform input distribution. Our code is to operate on two different Gilbert-Elliott channels (see Example 8 in Appendix E). Indeed, the capacity-achieving input distribution for these symmetric channels is uniform. The Gilbert-Elliott channels have the following parameters:

1) Channel GE-one:
   - crossover probability in good state: 0.01;
   - crossover probability in bad state: 0.175;
   - transition probability good state to bad state: 0.40;
   - transition probability bad state to good state: 0.40.
2) Channel GE-two:
   - crossover probability in good state: 0.01;
   - crossover probability in bad state: 0.170;
   - transition probability good state to bad state: 0.60;
   - transition probability bad state to good state: 0.55.

For these channels, we have the following bounds:

- $H(X_0|Y_{-L_0}^{L_0}, X_{-L_0}^{-1}) \leq 0.45$,
- $\epsilon_{L_0} \leq 7.39 \cdot 10^{-8}$,
- $\psi_{M_0-2} \approx \phi_{M_0-2} \approx 1$ (recall that these parameters approach 1 exponentially fast with M_0, which we have set to 40),
- $\psi_0 \leq 2.1$.

From these bounds, in step 1b of Section VII one may compute: $\varepsilon_1 = 2.36 \cdot 10^{-6}$, $\varepsilon_2 \approx 0$, and $\varepsilon_3 = 4.73 \cdot 10^{-6}$.

We designed a set of universal polar codes using the procedure in Section VII, with rates from 0.15 to 0.30. Specifically, in line with Remark 10, we have taken ψ_0 as 1 in step 2b. The simulation results of these of codes on both Gilbert-Elliott channels are shown in Figure 11. The decoding of these codes was done using a list-decoder [25]. Indeed, utilizing a larger list size for decoding improves the word error rate considerably.

Observe that the codes may operate on additional channels that satisfy the above bounds. Two examples are a BSC with crossover probability 0.09, having capacity 0.44, and BEC with erasure probability 0.45, having capacity 0.45. Both channels are memoryless so they trivially satisfy the other bounds. Simulation results of the *same* codes for these channels, again using a list-decoder, are shown in Figure 12.

We wish to emphasize that the numerical results are far better than the upper bounds on decoding error (step 3 in the method of Section VII). Indeed, the upper bound (sum of Bhattacharyya parameters of non-frozen indices) for rate 0.15 is 1.9 and the upper bound for rate 0.27 is 3960.41. Both are non-informative and extremely pessimistic. When constructing the codes, as non-frozen indices we have selected those with the lowest upper bounds on the Bhattacharyya parameter.

IX. A Contraction Inequality

In this section we introduce a contraction inequality that will be useful in proving a sufficient condition for forgetfulness in Section X. 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 $d(\mathbf{x}^T\mathsf{M}, \mathbf{y}^T\mathsf{M}) \leq \tau(\mathsf{M})d(\mathbf{x}, \mathbf{y})$.

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 [32, Chapter 3.1], [33, Section 2] is defined for nonnegative vectors with the same support.

[7] In fact, since we are dealing with states, the effective alphabet for the construction algorithm is non-binary, and we should have referred to [30] and [31] as well.

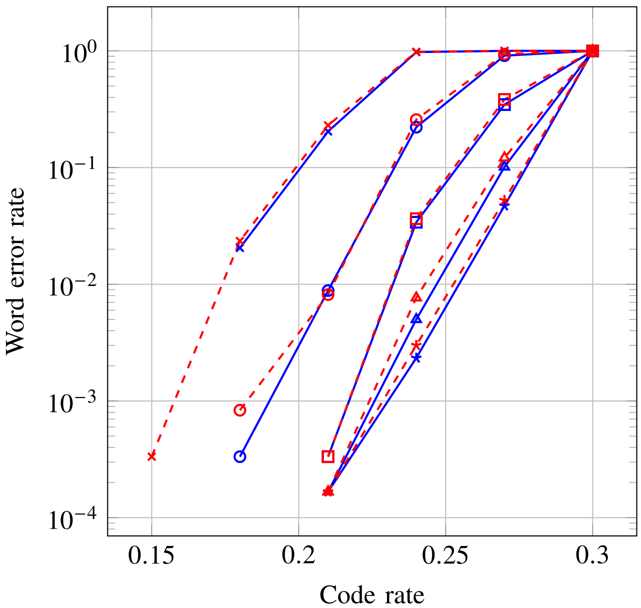

Fig. 11. Word error rate of a universal polar code of length 212992, used on two different Gilbert-Elliott channels: GE-one, in solid blue (—); and GE-two, in dashed red (- -). We use a list-decoder for the decoding, employing list sizes $L = 1$ (×), $L = 2$ (○), $L = 4$ (□), $L = 8$ (△), and $L = 16$ (⋆).

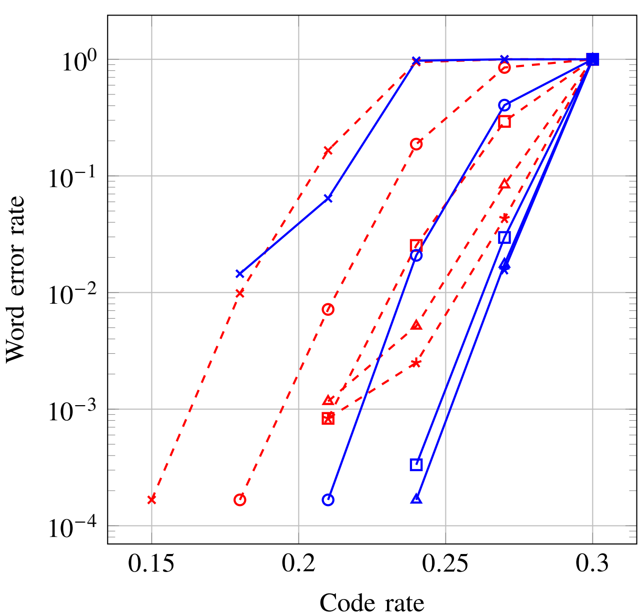

Fig. 12. Word error rate of the same universal polar code as in Figure 11, used on two additional channels: BEC, in solid blue (—); and BSC, in dashed red (- -). Again, we use a list-decoder for the decoding, employing list sizes $L = 1$ (×), $L = 2$ (○), $L = 4$ (□), $L = 8$ (△), and $L = 16$ (⋆).

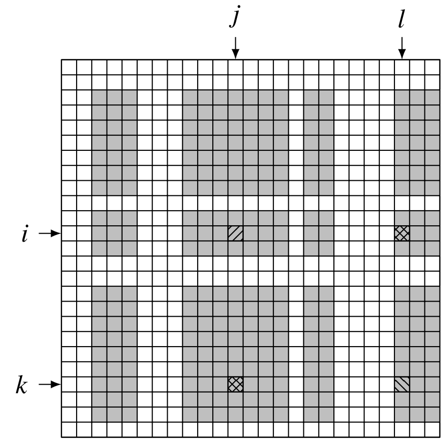

Fig. 13. An illustration of a subrectangular matrix. Each of the small squares is an element of the matrix. The white squares contain zeros, whereas the filled squares contain positive values. Elements $(\mathsf{M})_{i,j}$ and $(\mathsf{M})_{k,l}$, denoted with diagonal lines (▨ and ▧ respectively), 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 (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

$$d(\mathbf{x}, \mathbf{y}) \triangleq \max_{j,l \in \sigma(\mathbf{x})} \ln \frac{x_j / y_j}{x_l / y_l} = \ln \max_{j,l \in \sigma(\mathbf{x})} \frac{x_j / y_j}{x_l / y_l}. \tag{132}$$

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 [32, 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.

Definition 14 (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 13. 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 32. *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:*

$$(\mathsf{M})_{i,j} = 0 \iff i \notin \mathcal{N}_{\mathrm{r}}(\mathsf{M}) \text{ or } j \notin \mathcal{N}_{\mathrm{c}}(\mathsf{M}). \tag{133}$$

3) *The matrix* M *satisfies*

$$(\mathsf{M})_{i,j} \neq 0 \iff i \in \mathcal{N}_{\mathrm{r}}(\mathsf{M}) \text{ and } j \in \mathcal{N}_{\mathrm{c}}(\mathsf{M}). \tag{134}$$

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 (133) 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 (134) 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 (134), $i, k \in \mathcal{N}_{\mathrm{r}}(\mathsf{M})$ and $j, l \in \mathcal{N}_{\mathrm{c}}(\mathsf{M})$. Then, a second of use of (134) implies that $(\mathsf{M})_{i,l} \neq 0$ and $(\mathsf{M})_{k,j} \neq 0$. Therefore, M is subrectangular. ■

Observe from (133) 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, (134) 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 33. *If* M *is a nonzero subrectangular matrix and* $\mathbf{x}, \mathbf{y}$ *are nonnegative vectors such that* $\left\|\mathbf{x}^T\mathsf{M}\right\|_1 > 0$ *and* $\left\|\mathbf{y}^T\mathsf{M}\right\|_1 > 0$, *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 (133), 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 34. *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 33, every row of TM is either all-zeros, or has the same support as the other nonzero rows of TM. This implies, by (134), that TM is subrectangular. ■

We remark that a converse to Corollary 34 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 (134), if M is subrectangular and $(i, j) \in \sigma(\mathsf{M})$ then $(\mathsf{M})_{i,j} > 0$.

Definition 15 (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

$$\tau(\mathsf{M}) \triangleq \sup_{\mathbf{x}>0,\, \mathbf{y}>0} \frac{d(\mathbf{x}^T\mathsf{M}, \mathbf{y}^T\mathsf{M})}{d(\mathbf{x}, \mathbf{y})}. \tag{135}$$

- If M is the zero matrix, then $\tau(\mathsf{M}) = 0$.
- If M is not subrectangular, then $\tau(\mathsf{M}) = 1$.

By Lemma 33 and the positivity of $\mathbf{x}$ and $\mathbf{y}$, the numerator of (135) is well-defined. That is, $\mathbf{x}^T\mathsf{M}$ and $\mathbf{y}^T\mathsf{M}$ have the same support. The denominator of (135) 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 (135) is well-defined, we use the convention $0/0 = 0$. Observe that the supremum in (135) is obtained for $\mathbf{x} \neq c\mathbf{y}$ for $c > 0$.

The Birkhoff contraction coefficient [32, Chapter 3], [34] 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 13 and Lemma 33, 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

$$\tau(\mathsf{M}) = \tau(\mathsf{M}^T), \tag{136}$$

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 [32, Section 3.4] (see [34, Theorem 1.1] for an alternative proof).

Theorem 35. *If* M *is subrectangular and nonzero, then*

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

where

$$\phi(\mathsf{M}) \triangleq \min_{\substack{i,k\in\mathcal{N}_\mathrm{r}(\mathsf{M}),\\ j,l\in\mathcal{N}_\mathrm{c}(\mathsf{M})}} \frac{(\mathsf{M})_{i,j}(\mathsf{M})_{k,l}}{(\mathsf{M})_{i,l}(\mathsf{M})_{k,j}} > 0. \tag{137}$$

Since M is subrectangular and nonzero, all index pairs on the right-hand side of (137) 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 [32, Section 3.4] and [34, Theorem 1.1] assume no all-zero columns in M. However, as explained after Definition 15, they hold without change for our slightly generalized definition of the Birkhoff contraction coefficient.

By Definition 15 and Theorem 35, 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 36. *If* $\mathbf{x}, \mathbf{y}$ *are nonnegative vectors such that* $\sigma(\mathbf{x}) = \sigma(\mathbf{y})$ *and* M *is subrectangular, then*

$$d(\mathbf{x}^T\mathsf{M}, \mathbf{y}^T\mathsf{M}) \leq \tau(\mathsf{M})d(\mathbf{x}, \mathbf{y}). \tag{138}$$

Proof: The claim is trivial if $\mathbf{x} = \mathbf{y} = \mathbf{0}$. If $\mathbf{x}, \mathbf{y}$ are positive, the claim follows from Definition 15 and Theorem 35. 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,

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

Finally, observe that $(1-\sqrt{x})/(1+\sqrt{x})$ is a decreasing function of x when $x \geq 0$; 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, $\phi(\tilde{\mathsf{M}}) \geq \phi(\mathsf{M})$. Thus, we must have $\tau(\tilde{\mathsf{M}}) \leq \tau(\mathsf{M})$, which completes the proof. ■

In the following lemma we prove an inequality, adapted from the proof of [33, Lemma 5], that is useful in the sequel.

Lemma 37. *Let* $\alpha_i > 0$*,* $\beta_i > 0$*, and* $\gamma_i \geq 0$ *for all* i*. Assume that* $\gamma_i > 0$ *for some* i*. Then,*

$$\min_i \frac{\alpha_i}{\beta_i} \leq \frac{\sum_i \gamma_i\alpha_i}{\sum_i \gamma_i\beta_i} \leq \max_i \frac{\alpha_i}{\beta_i}. \tag{139}$$

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 $\theta_i \geq 0$ for all i and $\sum_i \theta_i = 1$. That is, the ratio on the left-hand side is a convex combination of the ratios 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 38. *Let* M *be a subrectangular matrix and let* T *be a nonnegative matrix. Then,*

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

If, in addition, T *is subrectangular then*

$$\tau(\mathsf{TM}) \leq \tau(\mathsf{T})\tau(\mathsf{M}). \tag{140}$$

Remark 11. Two remarks are in order. First, we note that (140) is adapted from [32, 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 (136), $\tau(\mathsf{TM}) = \tau((\mathsf{TM})^T) = \tau(\mathsf{M}^T\mathsf{T}^T)$ 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 34, 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 (137); 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 (134), $(i_0, j_0), (k_0, l_0) \in \sigma(\mathsf{TM})$. 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,

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

where (a) is by the left-hand inequality of (139), 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 $\phi(\mathsf{TM}) \geq \phi(\mathsf{M})$ and, since $(1 - \sqrt{x})/(1 + \sqrt{x})$ is a decreasing function of x for $x \geq 0$ (see the proof of Corollary 36), we conclude that $\tau(\mathsf{TM}) \leq \tau(\mathsf{M})$.

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 (138) yield

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

Thus, by (135), $\tau(\mathsf{TM}) \leq \tau(\mathsf{T})\tau(\mathsf{M})$. ■

Applying Lemma 38 to a product of m subrectangular matrices $\mathsf{M}_1, \mathsf{M}_2, \ldots, \mathsf{M}_m$, we obtain

$$\tau(\mathsf{M}_1\mathsf{M}_2\cdots\mathsf{M}_m) \le \prod_{\ell=1}^{m} \tau(\mathsf{M}_\ell). \tag{141}$$

Corollary 36 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 39. *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* $\|\mathbf{x}^T\mathsf{T}\mathsf{M}\|_1 > 0$ *and* $\|\mathbf{y}^T\mathsf{T}\mathsf{M}\|_1 > 0$,

$$d(\mathbf{x}^T\mathsf{T}\mathsf{M}, \mathbf{y}^T\mathsf{T}\mathsf{M}) \le 4\ln\left(\frac{1+\tau(\mathsf{M})}{1-\tau(\mathsf{M})}\right). \tag{142}$$

Since M is subrectangular, $\tau(\mathsf{M}) < 1$, which implies that the right-hand side of (142) is finite.

Proof: There is nothing to prove if $\mathsf{T}\mathsf{M} = 0$, so we assume that $\mathsf{T}\mathsf{M}$ is nonzero. By Corollary 34, $\tilde{\mathsf{M}} = \mathsf{T}\mathsf{M}$ 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 33, and subrectangularity of $\tilde{\mathsf{M}}$,

$$\sigma(\mathbf{e}_{i_0}^T\tilde{\mathsf{M}}) = \sigma(\mathbf{x}^T\tilde{\mathsf{M}}) = \mathcal{N}_{\mathrm{c}}(\tilde{\mathsf{M}}). \tag{143}$$

By the symmetry and triangle inequality properties of the projective distance [32, 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 38 and since $\ln((1+x)/(1-x))$ is monotone increasing for $0 \le x < 1$, (142) 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 (137). The right-hand equality follows directly from Theorem 35, so we concentrate on proving the inequality.

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

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

The denominator is positive by (143), so ρ_j is well-defined. Now, for $j, l \in \mathcal{N}_{\mathrm{c}}(\tilde{\mathsf{M}})$,

$$\begin{aligned}\frac{\rho_j}{\rho_l} &= \frac{\sum\limits_{k\in\mathcal{N}_{\mathrm{r}}(\tilde{\mathsf{M}})} x_k(\tilde{\mathsf{M}})_{k,l}}{\sum\limits_{k\in\mathcal{N}_{\mathrm{r}}(\tilde{\mathsf{M}})} x_k(\tilde{\mathsf{M}})_{k,j}} \cdot \frac{(\tilde{\mathsf{M}})_{i_0,j}}{(\tilde{\mathsf{M}})_{i_0,l}} \\ &\overset{(a)}{\le} \max_{k\in\mathcal{N}_{\mathrm{r}}(\tilde{\mathsf{M}})} \frac{(\tilde{\mathsf{M}})_{k,l}}{(\tilde{\mathsf{M}})_{k,j}} \cdot \frac{(\tilde{\mathsf{M}})_{i_0,j}}{(\tilde{\mathsf{M}})_{i_0,l}} \\ &\overset{(b)}{\le} \max_{k\in\mathcal{N}_{\mathrm{r}}(\tilde{\mathsf{M}})} \frac{(\tilde{\mathsf{M}})_{k,l}}{(\tilde{\mathsf{M}})_{k,j}} \cdot \max_{i\in\mathcal{N}_{\mathrm{r}}(\tilde{\mathsf{M}})} \frac{(\tilde{\mathsf{M}})_{i,j}}{(\tilde{\mathsf{M}})_{i,l}},\end{aligned} \tag{144}$$

where (a) is by Lemma 37 and in (b) we maximize over a set that contains i_0.

Hence, recalling the definition of the projective distance, (132),

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

where (a) is by (144) and (b) follows from the definition of ϕ in (137). This completes the proof. ■

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

Proposition 40. *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 $\|\mathbf{x}^T\mathsf{M}_1^m\mathsf{T}\|_1 > 0$ *and* $\|\mathbf{y}^T\mathsf{M}_1^m\mathsf{T}\|_1 > 0$, *then*

$$\ln\left(\frac{\|\mathbf{x}^T\mathsf{M}_1^m\mathsf{T}\|_1}{\|\mathbf{y}^T\mathsf{M}_1^m\mathsf{T}\|_1} \cdot \frac{\|\mathbf{y}^T\mathsf{M}_1^m\|_1}{\|\mathbf{x}^T\mathsf{M}_1^m\|_1}\right) \le d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \cdot \prod_{\ell=2}^{m} \tau(\mathsf{M}_\ell). \tag{145}$$

Proof: Since $\|\mathbf{x}^T\mathsf{M}_1^m\mathsf{T}\|_1 > 0$, 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 (145) is well-defined. We will show that

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

The right-hand side is well-defined since, by Corollary 34, M_r^s is subrectangular for any $1 \le r \le s \le m$ and by Lemma 33. Then, as $\sigma(\tilde{\mathbf{x}}) = \sigma(\tilde{\mathbf{y}})$ by Lemma 33, (145) will follow from Corollary 36 and (141).

Denote $J = \sigma(\tilde{\mathbf{x}}^T\mathsf{M}_2^m) = \sigma(\tilde{\mathbf{y}}^T\mathsf{M}_2^m) = \mathcal{N}_{\mathrm{c}}(\mathsf{M}_2^m)$, where the equalities are by Lemma 33 and subrectangularity. By the right-hand inequality of (139),

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

Next, denote by $t_j = \|(\mathsf{T})_{j,:}\|_1$ 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 (139) yields

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

Combining, we obtain

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

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 41. *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

$$\ln \left(\frac{\left\| \mathbf{x}^T \mathsf{R} \mathsf{T} \right\|_1}{\left\| \mathbf{y}^T \mathsf{R} \mathsf{T} \right\|_1} \cdot \frac{\left\| \mathbf{y}^T \mathsf{R} \right\|_1}{\left\| \mathbf{x}^T \mathsf{R} \right\|_1} \right) \leq 4 \ln \left(\frac{1 + \tau(\mathsf{M}_1)}{1 - \tau(\mathsf{M}_1)} \right) \cdot \prod_{\ell=2}^{m} \tau(\mathsf{M}_\ell). \quad (146)$$

Proof: The claim follows from Corollary 34, Lemmas 38 and 39, and Proposition 40. ■

Observe that (146) remains true if we replace 'ln' with 'log'.

Discussion. Our Proposition 40 and Corollary 41 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 $a \in \mathcal{A}$,[8] 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.

X. 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 IX 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.

[8]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.

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,

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

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., [22, Proposition 4.1]). This implies [22, 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, $\mathbb{P}\left(B_n = b\right) = \mathbb{P}\left(A_n \in f^{-1}(b)\right)$. 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.

Definition 16 (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

$$(\mathsf{M}(b))_{i,j} = \begin{cases} p(j|i), & \text{if } f(j) = b \\ 0, & \text{otherwise.} \end{cases} \quad (147)$$

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,

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

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, $(\boldsymbol{\pi})_i = \mathbb{P}\left(A_0 = i\right)$ and

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

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

$$\begin{aligned}\mathbb{P}\left(B_1^n = b_1^n\right) &= \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\right\|_1 \\ &= \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1)\mathsf{M}(b_2) \cdots \mathsf{M}(b_n)\right\|_1.\end{aligned} \tag{148}$$

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

$$\begin{aligned}\mathbb{P}\left(A_{n+1} = a, B_1^n = b_1^n\right) &= (\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\mathsf{M})_a \\ &= \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\mathsf{M}\mathbb{I}_a\right\|_1 \\ &= \left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\mathsf{T}_a\right\|_1,\end{aligned} \tag{149}$$

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 (148) and (149),

$$\begin{aligned}\mathbb{P}\left(A_{n+1} = a | B_1^n = b_1^n\right) &= \frac{\mathbb{P}\left(A_{n+1} = a, B_1^n = b_1^n\right)}{\mathbb{P}\left(B_1^n = b_1^n\right)} \\ &= \frac{\left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\mathsf{T}_a\right\|_1}{\left\|\boldsymbol{\pi}^T \mathsf{M}(b_1^n)\right\|_1}.\end{aligned} \tag{150}$$

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 (150) must also be positive.

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

$$\mathbb{P}\left(B_1^n = b_1^n | A_0 = a_0\right) = \left\|\mathbf{e}_{a_0}^T \mathsf{M}(b_1^n)\right\|_1, \tag{151}$$

$$\mathbb{P}\left(A_{n+1} = a | B_1^n = b_1^n, A_0 = a_0\right) = \frac{\left\|\mathbf{e}_{a_0}^T \mathsf{M}(b_1^n)\mathsf{T}_a\right\|_1}{\left\|\mathbf{e}_{a_0}^T \mathsf{M}(b_1^n)\right\|_1}, \tag{152}$$

provided that the probability in (151) is positive.

In (148)–(152), 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 (150) and (152). They are well-defined with probability 1. In other words, we can always compute their values via (150) and (152); with probability 0 will the denominators on the right-hand sides of these equations equal 0. These random variables are of interest because

$$I(A_0; A_{n+1} | B_1^n) = \mathbb{E}\left[\log \frac{\mathbb{P}\left(A_{n+1} | B_1^n, A_0\right)}{\mathbb{P}\left(A_{n+1} | B_1^n\right)}\right]. \tag{153}$$

Using (150) and (152) we write this as

$$\begin{aligned}&I(A_0; A_{n+1} | B_1^n) \\ &= \mathbb{E}\left[\log\left(\frac{\left\|\mathbf{e}_{A_0}^T \mathsf{M}(B_1^n)\mathsf{T}_{A_{n+1}}\right\|_1}{\left\|\boldsymbol{\pi}^T \mathsf{M}(B_1^n)\mathsf{T}_{A_{n+1}}\right\|_1} \cdot \frac{\left\|\boldsymbol{\pi}^T \mathsf{M}(B_1^n)\right\|_1}{\left\|\mathbf{e}_{A_0}^T \mathsf{M}(B_1^n)\right\|_1}\right)\right].\end{aligned} \tag{154}$$

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 [22, Theorem 4.3], [35, Theorem 8.9], which harks back to the days of A. A. Markov [36], 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.[9]

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 (146), which bounds expressions of a form similar to the argument of the expectation in (154).

Remark 12. An expression similar to (153) 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

[9]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.

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.[10] 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 (146) 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.

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 13. 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 [37], [38]. 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 42. *Suppose the HMM (A_n, B_n) satisfies Condition K. Then, for every $\epsilon > 0$ there exists an integer λ such that if $n \ge \lambda$ then*

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

The proof is given in the next subsection, and will follow from Proposition 47, 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 41 in (154) 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 $\mathsf{M}^{k_0} > 0$ [32, Theorem 1.4]. Therefore, by Theorem 42, 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., [39].

Corollary 43. *Suppose the HMM (A_n, B_n) satisfies Condition K. Then, for every $\epsilon > 0$ there exists an integer λ such that if $n \ge \lambda$ then*

$$I(A_1; A_n|B_1^n) \le \epsilon. \tag{155}$$

and

$$I(A_0; A_n|B_1^n) \le \epsilon. \tag{156}$$

Proof: The conditions of Theorem 42 are satisfied. Let λ be such that $I(A_1; A_n|B_2^{n-1}) \le \epsilon$ for any $n \ge \lambda$.

We first show (155). Recall that B_j is a function of A_j for any j. Thus, for any $n \ge \lambda$, $I(A_1, B_1; A_n, B_n|B_2^{n-1}) = I(A_1; A_n|B_2^{n-1}) \le \epsilon$. Therefore,

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

Since mutual information is nonnegative, each of the summands on the right-hand side is upper-bounded by ϵ. This yields (155).

To see (156), since $A_0 \multimap (A_1, B_1^n) \multimap A_n$, we use (2) and obtain

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

as required. ■

We remark that under the same conditions as Corollary 43 we also obtain $I(A_1; B_n|B_1^{n-1}) \le \epsilon$ and $I(A_0; B_n|B_1^{n-1}) \le \epsilon$.

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.[11] Suppose that both HMMs satisfy Condition K. Then, by Corollary 43, for every $\epsilon > 0$ there exists an integer λ such that if $n \ge \lambda$ then $I(A_1; A_n|B_1^n) \le \epsilon$ and $I(A_1; A_n|C_1^n) \le \epsilon$.

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 (155) holds with $B_n = (X_n, Y_n)$ or $B_n = Y_n$ for any n. In

[10] 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.

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

particular, for any $\epsilon > 0$ there exists an integer λ such that for any $k \geq \lambda$ we have

$$I(S_1; S_k | X_1^k, Y_1^k) \leq \epsilon,$$
$$I(S_1; S_k | Y_1^k) \leq \epsilon.$$

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

C. Proof of Theorem 42

The goal of this subsection is to prove Theorem 42. To this end, we make the following definition.

Definition 17 ($(n_\star, \delta_\star, \tau_\star)$-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

$$\mathbb{P}\left(\tau(\mathsf{M}(B_1^{n_\star})) \leq \tau_\star | A_0 = a_0\right) \geq 1 - \delta_\star, \quad \forall a_0 \in \mathcal{A}. \quad (157)$$

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 $(n_\star, \delta_\star, \tau_\star)$.

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

In Lemma 44, 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 satisfies 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 44. *If the HMM (A_n, B_n) satisfies Condition K then there exist a positive integer $n_\star$ and constants $\delta_\star < 1$ and $0 \leq \tau_\star < 1$ such that (157) is satisfied.*

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

1) For any $i, j \in \mathcal{A}$, $(\mathsf{M}^{k_0})_{i,j} \geq \gamma_0$. This follows from M being aperiodic and irreducible, so some power of it must be strictly positive [32, 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, $0 \leq \tau_\star < 1$.

Denote by $\mathcal{A}'$ the set of states that can lead to $f^{-1}(\beta_1)$ and then emit the observation sequence $\beta_1^{l_0}$,

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

That is, there is positive probability that the next l_0 observations after any state in $\mathcal{A}'$ is 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 (157) is satisfied with $n_\star = k_0 + l_0$ and $\delta_\star = 1 - \alpha_0 \gamma_0 < 1$. Indeed, for any $a_0 \in \mathcal{A}$,

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

where (a) is by Corollary 34 and Lemma 38, (b) is by Condition K, (c) is by the Markov property, and (d) is by (151). ∎

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

$$N_1(\tau) = \min\{n : \tau(\mathsf{M}(B_1^n)) \leq \tau\},$$
$$N_{k+1}(\tau) = \min\{n : \tau(\mathsf{M}(B_{N_k+1}^{N_k+n})) \leq \tau\}, \quad k \geq 1.$$

That is, the random variable $N_1(\tau)$ is the time of the first occurrence of a sequence that induces a Birkhoff contraction coefficient 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 45. *Let (A_n, B_n) be an $(n_\star, \delta_\star, \tau_\star)$-KHMM. If $\delta_\star > 0$, there exist $\gamma > 0$ and $0 \leq \rho < 1$ such that for any positive integer n_1,*

$$\mathbb{P}\left(N_1(\tau_\star) \geq n_1 \mid A_0 = a_0\right) \leq \gamma \rho^{n_1}, \quad \forall a_0 \in \mathcal{A}. \quad (158)$$

Proof: Let $T_0 = 1$ and denote, for any positive integer k, the random variable $T_k = \tau(\mathsf{M}(B_1^{k n_\star}))$. Observe that, by (157), $\mathbb{P}\left(T_1 \leq \tau_\star | A_0 = a_0\right) \geq 1 - \delta_\star$ for any $a_0 \in \mathcal{A}$.

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

$$\mathbb{P}\left(T_k > \tau_\star \;\middle|\; T_{k-1} > \tau_\star, A_0 = a_0\right) \leq \delta_\star. \quad (159)$$

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

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

where (a) is because, by Lemma 38, 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 (157). Rearranging yields (159). We remark that (159) is also true without conditioning on $\{T_{k-1} > \tau_\star\}$.

Thus,

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

where (a) is by Lemma 38, by which the second summand in the first equality must be 0, and (b) is by (159). We conclude that for any integer k and any $a_0 \in \mathcal{A}$,

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

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 (158) 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 45 because this is the more interesting case. Clearly, Lemma 45 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 42.

The upper bound in (158) 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 46. *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 $n_1, n_2, \ldots, n_m$,*

$$\begin{aligned}
&\mathbb{P}\left(N_1(\tau_\star) \ge n_1, N_2(\tau_\star) \ge n_2, \ldots, N_m(\tau_\star) \ge n_m\right) \\
&\qquad\qquad \le \gamma^m \rho^{n_1 + n_2 + \cdots + n_m}. \qquad (160)
\end{aligned}$$

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

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

(160) 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

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

where (a) is by definition of N_k and (b) is by (158). ∎

Proposition 47. *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

$$I(A_0; A_{n+1} | B_1^n) \le 4 \log\left(\frac{1+\tau_\star}{1-\tau_\star}\right) \tau_\star^m + \alpha \frac{(\gamma n)^m}{m!} \rho^{n+1}. \qquad (161)$$

Proof: Observe that the right-hand side of (160) 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 [40, p. 38], it is shown that $p(n,m) = \binom{n-1}{m-1}$. Thus, by (160),

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

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,

$$\begin{aligned}\mathbb{P}\left(D_n \le m\right) &= \mathbb{P}\left(D_n < m+1\right)\\ &= \mathbb{P}\left(\sum_{k=1}^{m+1} N_k(\tau_\star) \ge n+1\right)\\ &\le \gamma \frac{(n\gamma)^m}{m!}\rho^{n+1}. \end{aligned} \tag{162}$$

Recall from (154) 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

$$0 \le I(A_0; A_{n+1}|B_1^n = b_1^n) = \mathbb{E}\left[J|B_1^n = b_1^n\right] \le \log|\mathcal{A}|, \tag{163}$$

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 $m \le n$,

$$\begin{aligned}&\mathbb{E}\left[J|D_n > m\right]\mathbb{P}\left(D_n > m\right)\\ &\quad= \sum_{b_1^n : D_n > m} \mathbb{E}\left[J|B_1^n = b_1^n\right]\mathbb{P}\left(B_1^n = b_1^n\right)\\ &\quad\overset{\text{(a)}}{=} \sum_{b_1^n : D_n \ge m+1} \mathbb{E}\left[J|B_1^n = b_1^n\right]\mathbb{P}\left(B_1^n = b_1^n\right)\\ &\quad\overset{\text{(b)}}{\le} 4\log\left(\frac{1+\tau_\star}{1-\tau_\star}\right)\cdot\tau_\star^m, \end{aligned} \tag{164}$$

where (a) is because D_n is integer valued and (b) is by Lemma 38 and Corollary 41. Moreover,

$$\begin{aligned}&\mathbb{E}\left[J|D_n \le m\right]\mathbb{P}\left(D_n \le m\right)\\ &\quad= \sum_{b_1^n : D_n \le m} \mathbb{E}\left[J|B_1^n = b_1^n\right]\mathbb{P}\left(B_1^n = b_1^n\right)\\ &\quad\overset{\text{(a)}}{\le} \log|\mathcal{A}|\cdot\mathbb{P}\left(D_n \le m\right)\\ &\quad\overset{\text{(b)}}{\le} \log|\mathcal{A}|\cdot\gamma\frac{(n\gamma)^m}{m!}\rho^{n+1}, \end{aligned} \tag{165}$$

where (a) is by the right-hand inequality of (163) and (b) is by (162).

Thus, for any $m \le n$ we have by (164) and (165),

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

Denoting $\alpha = \gamma\cdot\log|\mathcal{A}|$ completes the proof. ■

Remark 14. We note in passing that, if desired, one can set $m = \theta n$ in (161) 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 [40, p. 52]. We set $m = \theta n$, and upper-bound the second summand in the right-hand side of (161) 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 $\lim_{\theta\to 0}(\gamma e/\theta)^\theta = 1$,[12] 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 (161). A different strategy is outlined in the proof of Theorem 42.

We are now ready to prove Theorem 42.

Proof of Theorem 42: By Lemma 44, (A_n, B_n) is an $(n_\star, \delta_\star, \tau_\star)$-KHMM for some $n_\star$, $\delta_\star$, $\tau_\star$. Let

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

Case 1: If $\delta_\star = 0$ then at time $\lambda = (m+1)n_\star$ the sequence B_1^λ can be divided into $m+1$ contiguous sequences of length $n_\star$, each inducing a Birkhoff contraction coefficient less than $\tau_\star$. Therefore, using Corollary 41 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 47. 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 (161) is upper-bounded by ϵ. ■

Discussion. The upper bound in Proposition 47 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 38, 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 42. A more refined analysis, that considers a KHMM for which multiple combinations $(n_\star, \delta_\star, \tau_\star)$ are known may yield better bounds.

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

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 47.

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 48 ([3], [6], [41])**.** *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*

$$Z_{n+1} \le \kappa \cdot \begin{cases} Z_n^2, & B_{n+1} = 0, \\ Z_n, & B_{n+1} = 1, \end{cases} \qquad n \ge 0, \tag{166}$$

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

$$\lim_{n\to\infty} \mathbb{P}\left(Z_n \le 2^{-2^{n\beta}}\right) = \mathbb{P}\left(Z_\infty = 0\right). \tag{167}$$

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[13] for the general case can be found in [41]. Our fast polarization result is based on the proof of [41].

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 (166) for the Bhattacharyya parameter with $\kappa = 2$ was established in [2, Proposition 5]. Under memory, (166) was shown in [13, Theorem 2], with

$$\kappa = 2\psi_0, \tag{168}$$

where ψ_0 is a mixing parameter of the process; mixing parameters are defined in Lemma 7. Thus, the Bhattacharyya parameter polarizes fast to 0 with or without memory.

[13] In which (167) is replaced with $\lim_{n_0\to\infty} \mathbb{P}(\forall n \ge n_0,\ Z_n \le 2^{-2^{n\beta}}) = \mathbb{P}(Z_\infty = 0)$.

The proof in [41] establishes (167) 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

$$\mathbb{P}\left(Z_n \le \epsilon_a\right) \ge \mathbb{P}\left(Z_\infty = 0\right) - \delta_a, \quad \forall n \ge n_a. \tag{169}$$

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[14] — is universally smaller than η, which can be set as small as desired.[15] The ability to set η as small as desired is the key to obtaining *universal* fast polarization results. Namely, we prove the following proposition.

Proposition 49. *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 (166) 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*

$$\mathbb{P}\left(Z_n \le 2^{-2^{n\beta}} \text{ for all } n \ge n_0\right) \ge 1 - \delta. \tag{170}$$

Crucially, η and n_0 depend on the process Z_n only through κ. Inspection of the proof of [41] reveals that Proposition 49 will be true once it is shown that for any $\epsilon_a > 0$ and $\delta' > 0$ there exists n_a such that

$$\mathbb{P}\left(Z_n \le \epsilon_a \text{ for all } n \ge n_a\right) \ge 1 - \delta'. \tag{171}$$

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 51, which follows from Lemma 50, introduced and proved below.

Remark 15. Our statement of Proposition 49 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 50, is stated in a general manner, enabling its use for other kernels without change.

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

[15] 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 23, 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 (166). 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)].

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

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

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

$$\mu \triangleq \mathbb{E}\left[T\right] < 0. \tag{172}$$

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

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

and

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

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 [35, Theorem 2.1],

$$\mathbb{P}\left(\mathcal{A}_\alpha\right) = \lim_{n \to \infty} \mathbb{P}\left(\mathcal{A}_\alpha(n)\right). \tag{173}$$

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

$$\mathbb{P}\left(\mathcal{A}_\alpha\right) \leq e^{-r\alpha}. \tag{174}$$

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

$$\mathbb{P}\left(J_n < n(1-\gamma)\mu\right) \geq 1 - e^{-2n\left(\frac{\gamma\mu}{a+b}\right)^2}. \tag{175}$$

Since $\mu < 0$ by (172) and $0 < \gamma < 1$ by assumption, then $n(1-\gamma)\mu < 0$ in (175). We will see in Corollary 51 below that Lemma 50 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 $n \geq n_a$ and never (for any $n \in \mathbb{N}$) visits above α. This will be key to obtaining (171).

Proof: The proof combines two inequalities: (174) is essentially the Lundberg inequality [42, equation 15] and for (175) we call upon the Hoeffding inequality [43, Theorem 2]. Since the proof of the Lundberg inequality in [42] is for the continuous-time case, we provide a proof for the discrete-time case, adapted from the proof of [42].

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 [35, 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 [44, Theorem 9.3.3]. Observe that $g(0) = 1$ and, by [35, equation 21.23] and (172), $g'(0) = \mathbb{E}\left[T\right] < 0$. Thus, $g(s)$ is decreasing at $s = 0$, so $g(s) < 1$ for s small enough. On the other hand, by assumption on T,

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

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

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

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

$$r \triangleq \max_{s>0}\left\{s : \mathbb{E}\left[e^{sT}\right] = 1\right\}. \tag{176}$$

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,

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

where the last equality is by definition of r, (176). Define the (possibly infinite) stopping time

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

Then, by [45, Section 10.9], the stopped process

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

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 $\mathbb{P}(\mathcal{A}_\alpha(n)) = \mathbb{P}(\tau \leq n)$. Thus,

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

where (a) is because $\tilde{J}_{n\wedge\tau} \geq 0$, (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 $n \in \mathbb{N}$,

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

Thus, by (173),

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

This completes the proof of (174).

To prove (175), recall that by the Hoeffding inequality [43, Theorem 2], for any $t > 0$ we have

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

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

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

This completes the proof. ■

Corollary 51. *Under the same setting as in Lemma 50, for any $n_a \geq 0$, $\alpha > 0$, and $0 < \gamma < 1$ we have*

$$\begin{aligned}&\mathbb{P}\left(\{\forall n \geq n_a,\ J_n < n_a(1-\gamma)\mu\} \cap \mathcal{A}^{\mathrm{c}}_\alpha\right) \\ &\quad \geq 1 - \left(1 - e^{-2\left(\frac{\gamma\mu}{a+b}\right)^2}\right)^{-1} \cdot e^{-2n_a\left(\frac{\gamma\mu}{a+b}\right)^2} - e^{-r\alpha}. \qquad (177)\end{aligned}$$

Proof: Note that

$$\begin{aligned}&\mathbb{P}\left(\forall n \geq n_a,\ J_n < n_a(1-\gamma)\mu\right) \\ &\quad = \mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n_a(1-\gamma)\mu\}\right) \\ &\quad \geq \mathbb{P}\left(\bigcap_{n=n_a}^{\infty} \{J_n < n(1-\gamma)\mu\}\right) \\ &\quad = 1 - \mathbb{P}\left(\bigcup_{n=n_a}^{\infty} \{J_n \geq n(1-\gamma)\mu\}\right) \\ &\quad \overset{\text{(a)}}{\geq} 1 - \sum_{n=n_a}^{\infty} e^{-2n\left(\frac{\gamma\mu}{a+b}\right)^2} \\ &\quad = 1 - \left(\frac{1}{1 - e^{-2\left(\frac{\gamma\mu}{a+b}\right)^2}}\right) \cdot e^{-2n_a\left(\frac{\gamma\mu}{a+b}\right)^2}, \qquad (178)\end{aligned}$$

where (a) is by (175) and the union bound. Observing that

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

we obtain

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

Combining this inequality with (174) and (178) yields (177) and completes the proof. ■

Proof of Proposition 49: By inspection of the proof of [41], 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 $Z_0 \leq \eta$, then (171) 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

$$\bar{Z}_0 = \ln Z_0,$$

$$\bar{Z}_{n+1} = \begin{cases} 2\bar{Z}_n + \ln\kappa, & B_{n+1} = 0, \\ \bar{Z}_n + \ln\kappa, & B_{n+1} = 1, \end{cases} \quad n \geq 0. \qquad (179)$$

Then, by (166), $\ln Z_n \leq \bar{Z}_n$ for any n. Therefore, (171) will be true once we show that there exists n_a and η such that if $\bar{Z}_0 = \ln\eta$, then

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

Fix

$$0 < \zeta < 1/\kappa^2 \qquad (180)$$

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 (166),

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

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

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

where

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

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

Recall that $B_1, B_2, \ldots$ is a sequence of i.i.d. random variables with $\mathbb{P}(B_i = 0) = \mathbb{P}(B_i = 1) = 1/2$ for any i. Thus, $T_1, T_2, \ldots$ is a sequence of i.i.d. random variables. Denoting by T a random variable distributed according to their common distribution, we have $\mathbb{P}(T = \ln\kappa) = \mathbb{P}(T = \ln\kappa + \ln\zeta) = 1/2$. In particular, T is bounded:

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

Both a and b are positive by (180) and since $\kappa > 1$ by assumption. By definition, for any $\epsilon > 0$, $\mathbb{P}(T > a - \epsilon) \geq \mathbb{P}(T = a) = 1/2$. Moreover, by (180),

$$\mu = \mathbb{E}\left[T\right] = \frac{1}{2}\ln(\kappa^2\zeta) < 0.$$

Consequently, Corollary 51 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

$$\mathbb{E}\left[e^{rT}\right] = \frac{(\kappa\zeta)^r + \kappa^r}{2} = 1. \qquad (181)$$

Such r exists, as shown in the proof of Lemma 50. Denote for brevity

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

By Corollary 51, for any $0 < \gamma < 1$ and $n_a \geq 0$ we have

$$\begin{aligned}&\mathbb{P}\left(\{\forall n \geq n_a,\ J_n - J_0 < n_a(1-\gamma)\mu\} \cap \mathcal{A}^{\mathrm{c}}_{-J_0+\ln\zeta}\right) \\ &\quad \overset{\text{(a)}}{=} \mathbb{P}\left(\{\forall n \geq n_a,\ J_n < J_0 - n_a(1-\gamma)|\mu|\} \cap \mathcal{A}^{\mathrm{c}}_{-J_0+\ln\zeta}\right) \\ &\quad \geq 1 - (1 - e^{-2\gamma^2\theta^2})^{-1} e^{-2n_a\gamma^2\theta^2} - e^{-r(-J_0+\ln\zeta)}, \qquad (182)\end{aligned}$$

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 $\bar{Z}_n \leq J_n$ for any n. Hence,

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

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

Recall that $J_0 = \bar{Z}_0 = \ln\eta$. It remains to set η and n_a such that $\ln\eta < \ln\zeta$, $J_0 - n_a(1-\gamma)|\mu| \leq \ln\epsilon_a$, and the right-hand side of (182) 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 (182). We stress that these are arbitrary choices, and in practice should be optimized. We plug ζ into (181) and compute r, the largest positive solution of $\kappa^r + (2\kappa)^{-r} = 2$.

Next, we set J_0 so that $e^{-r(-J_0+\ln\zeta)} \leq \delta'/2$; one choice is $J_0 = \ln\zeta + \frac{1}{r}\ln(\delta'/2)$. Observe that indeed $J_0 = \ln\eta < \ln\zeta$ since $\delta' < 1$ (there is nothing to prove if $\delta' \geq 1$). 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 $J_0 - n_a|\mu|/2 \leq \ln\epsilon_a$ and $(1-e^{-2\gamma^2\theta^2})^{-1}e^{-2n_a\gamma^2\theta^2} \leq \delta'/2$. 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, $\mathbb{P}(\bar{Z}_n \leq \ln\epsilon_a \text{ for all } n \geq n_a) \geq 1-\delta'$. Thus, (171) 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 (166) 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 (179) 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 $P_{S_N} = P_{S_N}(s_N)$.

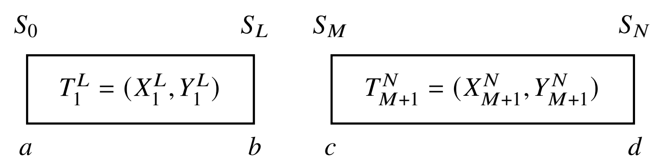

Fig. 14. Two blocks of a FAIM process, not necessarily of the same length. The state S_0, just before the first block, assumes value $a \in \mathcal{S}$. The final state of the first block, S_L, assumes value $b \in \mathcal{S}$. The state S_M, just before the second block, assumes value $c \in \mathcal{S}$. The final state of the second block, S_N, assumes value $d \in \mathcal{S}$.

Proof of Lemma 7: Although (34a) was already proved in [14, Lemma 5], we provide a proof here for completeness.

We will prove that (34) holds with

$$\psi_k = \begin{cases}\max_{s,\sigma} \dfrac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)\,\mathbb{P}(S_k = \sigma)}, & k > 0,\\ \max_{s} \dfrac{1}{\mathbb{P}(S_0 = s)}, & k = 0\end{cases} \tag{183}$$

and

$$\phi_k = \begin{cases}\min_{s,\sigma} \dfrac{\mathbb{P}(S_0 = s, S_k = \sigma)}{\mathbb{P}(S_0 = s)\,\mathbb{P}(S_k = \sigma)}, & k > 0,\\ 0, & k = 0.\end{cases} \tag{184}$$

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

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

To prove (34), 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 14). Then,

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

where (a) follows from the definition of ψ_k. This shows (34a). To see (34b) we follow the exact steps above up to just before inequality (a), and proceed with

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

Again, the inequality follows from the definition of ϕ_k.

For the case $M = L$, we need only establish (34a), as (34b) 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,

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

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 [22, 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 8: We will prove (37b). The proof of (37a) is identical, with the replacement of Y_a^b with X_a^b, Y_a^b throughout for any a and b.

The process (S_j, X_j, Y_j), $j \in \mathbb{Z}$ is FAIM, so it satisfies the Markov property (33). The proof follows from the following chain of inequalities.

$$\begin{aligned} I(S_1;S_\lambda|Y_1^\lambda) &\overset{(a)}{\geq} I(S_1;(Y_{\lambda+1}^k,S_m)|Y_1^\lambda) \\ &= I(S_1;Y_{\lambda+1}^k|Y_1^\lambda) + I(S_1;S_m|Y_1^k) \\ &\overset{(b)}{\geq} I(S_1;S_m|Y_1^k) \\ &\overset{(c)}{\geq} I(S_\ell;S_m|Y_1^k). \end{aligned}$$

We now justify the inequalities:

- (a) is by (2), noting that since $m \geq k \geq \lambda \geq 1$, (33) implies
$$S_1 \multimap (S_\lambda, Y_1^\lambda) \multimap (Y_{\lambda+1}^k, S_m);$$
- (b) is because mutual information is nonnegative;
- (c) is by (2), noting that since $\ell \leq 1$, (33) implies
$$S_m \multimap (S_1, Y_1^k) \multimap S_\ell$$
(observe that this is a Markov chain in reverse order of time).

This completes the proof. ∎

Proof of Lemma 9: The FAIM process is forgetful, so we let λ be the ϵ-recollection of the process. For this λ, (35) is satisfied.

By the chain rule for mutual information,

$$\begin{aligned} &I(S_0;S_{-k},S_k|X_{-\ell}^{-1},Y_{-\ell}^m) \\ &\quad= I(S_0;S_k|X_{-\ell}^{-1},Y_{-\ell}^m) + I(S_0;S_{-k}|X_{-\ell}^{-1},Y_{-\ell}^m,S_k). \end{aligned} \tag{185}$$

We will upper-bound each of the terms on the right-hand side of (185) 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:[16]

- (a) is by (35b) and stationarity.
- (b) is by (2), noting that (33) implies
$$S_0 \multimap (S_k, Y_0^k) \multimap (S_k, Y_{k+1}^m);$$
- (c) is because mutual information is nonnegative;
- (d) is by (2), noting that (33) implies
$$S_k \multimap (S_0, Y_0^m) \multimap (S_0, X_{-\ell}^{-1}, Y_{-\ell}^{-1})$$
(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.

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

The derivation for the second term in the right-hand side of (185) 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 (36a) and stationarity.
- (b) is by (2), noting that (33) 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 (33) 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 10: The FAIM process is forgetful, so we set λ as the ϵ-recollection of the process. The corollary holds for $k = 1$ by Lemma 9. We proceed by induction. Assume that the corollary 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,

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

where (a) is by the chain rule; (b) is by (2) and (33), used for the Markov chains (See Figure 15 for an illustration):

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

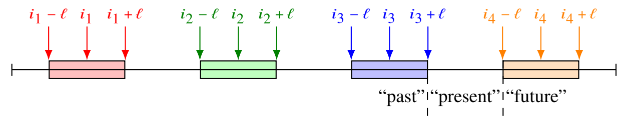

Fig. 15. Illustration of the timeline for $k = 4$. Given $S_{i_4-\ell}$, the "future" is independent of the "present" and "past." Given $S_{i_3+\ell}$, the "past" is independent of the "present" and "future."

which holds because $i_{k-1} \leq i_{k-1} + L_0 \leq i_k - L_0$ so $(S_{i_k-L_0}, S_{i_k+L_0}, C_{i_k})$ are independent of $S_{\mathbf{i}'}$ 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 $i_{k-1} + L_0 \leq i_k - L_0 \leq i_k$, so $(S_{\mathbf{i}'}, S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}, C_{\mathbf{i}'})$ are independent of S_{i_k} given $S_{i_{k-L_0}}$; finally, (c) is because $I(S_{\mathbf{i}'}; S_{\mathbf{i}'-L_0}, S_{\mathbf{i}'+L_0}|C_{\mathbf{i}'}) \leq (k-1) \cdot 2\epsilon$ by the induction hypothesis and $I(S_{i_k}; S_{i_k-L_0}, S_{i_k+L_0}|C_{i_k}) \leq 2\epsilon$ by Lemma 9. 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 15.

Lemma 52. *For any* $0 \leq x \leq 1/2$*,*

$$1 - h_2(x) \geq \frac{2}{\ln 2}\left(\frac{1}{2} - x\right)^2. \tag{186}$$

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 [46, Theorem 5.19], for any $x \in [\epsilon, 1/2]$, there exists $y \in [x, 1/2]$ such that

$$g(x) = \frac{2}{\ln 2}\left(\frac{1}{2} - x\right)^2 + \frac{g^{(4)}(y)}{4!}\left(\frac{1}{2} - x\right)^4.$$

However, $g^{(4)}(y) = 2(y^{-3} + (1-y)^{-3})/\ln 2 > 0$ for any $y \in [\epsilon, 1/2]$. Hence, $1 - h_2(x) \geq 2(1/2 - x)^2/(\ln 2)$ for any $0 \leq x \leq 1/2$ as well. ∎

Lemma 53. *For any* $0 \leq y \leq x \leq 1/2$*,*

$$h_2(x) - h_2(y) \geq \frac{1}{\ln 2}(x-y)(1-2y). \tag{187}$$

Proof: There is nothing to prove if $x = y$, so we assume that $y < x$. Due to the concavity of $h_2(x)$, for any $x_1 \leq x_2 \leq x_3$ we have

$$(x_3 - x_1)(h_2(x_2) - h_2(x_1)) \geq (x_2 - x_1)(h_2(x_3) - h_2(x_1)) \tag{188}$$

(see, for example, [47, Section 1.4.3], or [48, Exercise 6.17]). Setting $x_1 = y, x_2 = x, x_3 = 1/2$ in (188) we obtain

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

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

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

where the rightmost inequality is by (186). ∎

Lemma 54. *For any $x, y \in (0, 1/2)$, the function*

$$f(x, y) = h_2(h_2^{-1}(x) * h_2^{-1}(y)) - y \tag{189}$$

is increasing in x and decreasing in y.

Proof: Denote, for $x, y \in (0, 1/2)$,

$$g(x, y) = h_2(x * y) - h_2(y).$$

Then, $f(x, y) = g(h_2^{-1}(x), h_2^{-1}(y))$. The function $h_2(x)$ is monotone increasing over $[0, 1/2]$, so $h_2^{-1}(x)$ is also monotone increasing over $[0, 1/2]$. Therefore, the claim will be true once we establish that $g(x, y)$ is increasing in x and decreasing in y.

To this end, recall the function

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

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,

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

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,

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

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 15: It was shown in [6, Lemma 2.1] that

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

where

$$A = \sum_a p_a h_2(\alpha_a), \quad B = \sum_b q_b h_2(\beta_b).$$

Therefore,

$$\begin{aligned}
\sum_{a,b} p_a q_b \left(h_2(\alpha_a * \beta_b) - h_2(\beta_b)\right) &\geq h_2(h_2^{-1}(A) * h_2^{-1}(B)) - B \\
&= f(A, B),
\end{aligned}$$

where $f(\cdot, \cdot)$ was defined in (189). By (57), $A \geq \xi_1$ and $B \leq \xi_2$. Since, by Lemma 54, $f(A, B)$ 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,

$$\Delta(\xi_1, \xi_2) \triangleq h_2(h_2^{-1}(\xi_1) * h_2^{-1}(\xi_2)) - \xi_2. \tag{190}$$

It remains to show that $\Delta(\xi_1, \xi_2) > 0$.

To this end, observe that for any $x, y \in (0, 1/2)$,

$$\begin{aligned}
h_2(x * y) - h_2(y) &\overset{\text{(a)}}{\geq} \frac{1}{\ln 2}(x * y - y) \cdot (1 - 2y) \\
&= \frac{1}{\ln 2}x(1 - 2y)^2.
\end{aligned}$$

where (a) is by (187). Therefore,

$$\Delta(\xi_1, \xi_2) \geq \frac{1}{\ln 2}h_2^{-1}(\xi_1)\left(1 - 2h_2^{-1}(\xi_2)\right)^2 > 0. \qquad ∎$$

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 25: 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

$$p(f_0, g_0) \triangleq \mathbb{P}\left(F = f_0, G = g_0\right) = \sum_{\substack{a: f(a) = f_0, \\ g(a) = g_0}} p(a), \tag{191a}$$

$$q(f_0, g_0) \triangleq \mathbb{P}\left(\tilde{F} = f_0, \tilde{G} = g_0\right) = \sum_{\substack{a: f(a) = f_0, \\ g(a) = g_0}} q(a). \tag{191b}$$

With this notation we also have $p(g_0) = \mathbb{P}\left(G = g_0\right)$ and $p(f_0|g_0) = \mathbb{P}\left(F = f_0|G = g_0\right)$. The distributions $q(g_0), q(f_0|g_0)$ are similarly defined. By (83) and (191) we have for all $f_0 \in \{0, 1\}$ and $g_0 \in \mathcal{G}$,

$$\begin{aligned}
(1 - \varepsilon)q(f_0, g_0) \leq p(f_0, g_0) &\leq (1 + \varepsilon)q(f_0, g_0), \\
(1 - \varepsilon)q(g_0) \leq \quad p(g_0) \quad &\leq (1 + \varepsilon)q(g_0).
\end{aligned} \tag{192}$$

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 $|p(f_0|g_0) - q(f_0|g_0)| \le 3\varepsilon \cdot q(f_0|g_0)$. Thus, for any $g_0 \in \mathcal{G}$, since $\varepsilon < \frac{1}{6}$ by assumption,

$$d(g_0) \triangleq \sum_{f_0=0}^{1} |p(f_0|g_0) - q(f_0|g_0)| \le 3\varepsilon \sum_{f_0=0}^{1} q(f_0|g_0) = 3\varepsilon < \frac{1}{2}.$$

Since F and $\tilde{F}$ are binary, we conclude from [20, Theorem 17.3.3] that for any $g_0 \in \mathcal{G}$,

$$\begin{aligned} \left| H(F|G = g_0) - H(\tilde{F}|\tilde{G} = g_0) \right| &\le -d(g_0) \log \frac{d(g_0)}{2} \\ &\overset{(a)}{\le} -3\varepsilon \log \frac{3\varepsilon}{2}. \end{aligned} \tag{193}$$

Inequality (a) is true because $x \mapsto -x \log \frac{x}{2}$ is increasing for $0 \le x < \frac{2}{e} \approx 0.736$, and $0 \le d(g_0) \le 3\varepsilon < \frac{1}{2} < \frac{2}{e}$ by assumption.

Let $\sum^+$ denote summation over all $g_0 \in \mathcal{G}$ for which $p(g_0) \ge q(g_0)$, 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

$$\begin{aligned} \sum\nolimits^+ (p(g_0) - q(g_0)) &= -\sum\nolimits^- (p(g_0) - q(g_0)) \\ &= \frac{1}{2} \sum_{g_0} |p(g_0) - q(g_0)| \\ &\le \frac{\varepsilon}{2} \sum_{g_0} q(g_0) = \frac{\varepsilon}{2}, \end{aligned}$$

where the inequality is by (192). Hence, for any nonnegative function $h : \mathcal{G} \to \mathbb{R}^+$,

$$\begin{aligned} &\sum_{g_0} (p(g_0) - q(g_0)) h(g_0) \\ &\quad = \sum\nolimits^+ |p(g_0) - q(g_0)| h(g_0) - \sum\nolimits^- |p(g_0) - q(g_0)| h(g_0) \\ &\quad \le \left(\sup_{g_0} h(g_0) - \inf_{g_0} h(g_0) \right) \cdot \frac{1}{2} \sum_{g_0} |p(g_0) - q(g_0)| \\ &\quad \le \left(\sup_{g_0} h(g_0) - \inf_{g_0} h(g_0) \right) \cdot \frac{\varepsilon}{2}. \end{aligned} \tag{194}$$

Therefore,

$$\begin{aligned} &H(F|G) - H(\tilde{F}|\tilde{G}) \\ &= \sum_{g_0} p(g_0) H(F|G = g_0) - \sum_{g_0} q(g_0) H(\tilde{F}|\tilde{G} = g_0) \\ &\overset{(a)}{\le} \sum_{g_0} p(g_0) \left(H(\tilde{F}|\tilde{G} = g_0) - 3\varepsilon \log \frac{3\varepsilon}{2} \right) \\ &\quad - \sum_{g_0} q(g_0) H(\tilde{F}|\tilde{G} = g_0) \\ &= -3\varepsilon \log \frac{3\varepsilon}{2} + \sum_{g_0} (p(g_0) - q(g_0)) H(\tilde{F}|\tilde{G} = g_0) \\ &\overset{(b)}{\le} -3\varepsilon \log \frac{3\varepsilon}{2} + \left(\max_{g_0} H(\tilde{F}|\tilde{G} = g_0) - \min_{g_0} H(\tilde{F}|\tilde{G} = g_0) \right) \cdot \frac{\varepsilon}{2} \\ &\overset{(c)}{\le} \frac{\varepsilon}{2} - 3\varepsilon \log \frac{3\varepsilon}{2}, \end{aligned}$$

where (a) is by (193), (b) is by (194), and (c) is because the entropy of a binary random variable assumes values between 0 and 1.

Similarly,

$$\sum_{g_0} (p(g_0) - q(g_0)) h(g_0) \ge -\left(\sup_{g_0} h(g_0) - \inf_{g_0} h(g_0) \right) \cdot \frac{\varepsilon}{2},$$

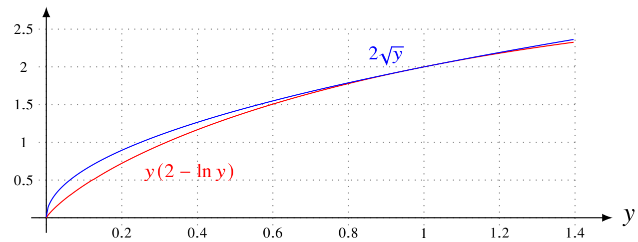

Fig. 16. Illustration of the inequality $y(2 - \ln y) \le 2\sqrt{y}$.

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 55 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 55. *For any $y > 0$, we have*

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

Proof: This inequality is illustrated in Figure 16. A formal proof follows. The Fenchel dual of $f(x) = e^x$ [49, 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 (33). However, in Section X, 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 X, 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

$$\mathbb{P}\left(B_n = b \,|\, A_n = j, B_1^{n-1}, A_1^{n-1} \right) = \mathbb{P}\left(B_n = b | A_n = j \right) = q(b|j). \tag{195}$$

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 (195), we have

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

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,[17] and whose transition probability function for any two states (j, b) and (i, k) is $\mathbb{P}\left(C_n = (j, b) \mid C_{n-1} = (i, k)\right) = p(j|i)q(b|j)$. 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 56. *Let A_n be a finite-state homogeneous Markov chain and let B_n be a probabilistic observation of A_n, as in (195). 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 [32, Theorem 1.4 and Section 4.2], also [22, Section 4.1].

Assume first that A_n is aperiodic and irreducible. Hence, there exists m such that $\mathbb{P}\left(A_n = j | A_{n-m} = i\right) > 0$ for all i, j, and n. Therefore, for any viable pairs (j, b) and (i, k),

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

Since the states of C_n consist only of viable pairs, we conclude that C_n is aperiodic and irreducible.

Next, assume that C_n is aperiodic and irreducible. Then, there exists m such that $\mathbb{P}\left(C_n = (j, b) | C_{n-m} = (i, k)\right) > 0$ 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),

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

Hence, A_n is aperiodic and irreducible. ■

Example 8. The Gilbert-Elliott channel [50] is a classic example of a channel with memory. It is defined as follows. The channel may be at one of two states, *good* and *bad*. In the good state, the channel is a binary symmetric channel (BSC) with crossover probability 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.

Assuming a symmetric channel input, we construct a deterministic 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.

[17] States for which $q(b|j) = 0$ can never appear with positive probability and are therefore removed.

[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," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 1994–2003, 2019.

[14] B. Shuval and I. Tal, "Fast polarization for processes with memory," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2004–2020, 2019.

[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] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, 2nd ed. Addison-Wesley, 1994.

[22] M. Iosifescu, *Finite Markov processes and their applications*, ser. Dover books on mathematics. Mineola, N.Y: Dover, 2007.

[23] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the lambert *W* function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 01 1996.

[24] 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.

[25] B. Shuval and I. Tal, "List decoding of universal polar codes," in *(ISIT'2020)*, Los Angeles, California, 2020.

[26] ——, "Fast polarization for processes with memory," *IEEE Trans. Inform. Theory*, vol. 65, no. 4, pp. 2004–2020, April 2019.

[27] K. E. Iverson, *A Programming Language*. New York: Wiley, 1962.

[28] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, School Comput. Commun. Sci., EPFL, Lausanne, Switzerland, 2009. [Online]. Available: https://infoscience.epfl.ch/record/138655/files/EPFL_TH4461.pdf

[29] I. Tal and A. Vardy, "How to construct polar codes," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6562–6582, October 2013.

[30] 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.

[31] O. Ordentlich and I. Tal, "An upgrading algorithm with optimal power law," *Accepted for publication in IEEE Transactions on Information Theory*, 2021.

[32] E. Seneta, *Non-negative matrices and Markov chains*, ser. Springer series in statistics. New York, NY: Springer, 2006.

[33] J. E. Cohen, "Contractive inhomogeneous products of non-negative matrices," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 86, no. 2, p. 351–364, 1979.

[34] 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.

[35] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1995.

[36] E. Seneta, "Markov and the birth of chain dependence theory," *International Statistical Review*, vol. 64, no. 3, pp. 255–263, 1996.

[37] 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.

[38] 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.

[39] 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.

[40] W. Feller, *An Introduction to Probability Theory and Its Applications volume 1*, 3rd ed. Wiley, 1968.

[41] I. Tal, "A simple proof of fast polarization," *IEEE Transactions on Information Theory*, vol. 63, no. 12, pp. 7617–7619, Dec 2017.

[42] H. U. Gerber, "Martingales in risk theory," *Bulletin / Association of Swiss Actuaries*, vol. 73, pp. 205–216, 1973.

[43] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[44] J. S. Rosenthal, *A first look at Rigorous Probability Theory*, 2nd ed. World Scientific, 2006.

[45] D. Williams, *Probability with Martingales*. Cambridge University Press, 1991.

[46] T. M. Apostol, *Mathematical analysis*, 2nd ed. Addison-Wesley, 1974.

[47] D. S. Mitrinović and P. M. Vasić, *Analytic inequalities*, ser. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1970.

[48] J. M. Steele, *The Cauchy-Schwarz Master Class*. Cambridge University Press, 2004.

[49] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[50] 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.