# Fast Polarization for Processes with Memory

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

***Abstract*—Fast polarization is crucial for the performance guarantees of polar codes. In the memoryless setting, the rate of polarization is known to be exponential in the square root of the block length. A complete characterization of the rate of polarization for models with memory has been missing. Namely, previous works have not addressed fast polarization of the high entropy set under memory. We consider polar codes for processes with memory that are characterized by an underlying ergodic finite-state Markov chain. We show that the rate of polarization for these processes is the same as in the memoryless setting, both for the high and for the low entropy sets.**

***Index Terms*—Polar codes, rate of polarization, fast polarization, channels with memory, Markov processes**

## I. Introduction

MEMORY is prevalent in many communication scenarios. Examples include finite-state channels (FSCs) such as intersymbol interference channels and correlated fading channels, and coding for input-constrained systems. In this research we show that polar codes can be used directly for a large class of scenarios with memory. This allows one to leverage the attractive properties of polar codes — such as low complexity encoding and decoding, explicit construction, and sound theoretical basis — for scenarios with memory.

A fundamental problem of information theory is estimating a block $X_1^N = (X_1, X_2, \ldots, X_N)$ from observations $Y_1^N = (Y_1, Y_2, \ldots, Y_N)$. In a channel-coding scenario, $X_1^N$ may be the input to a channel and $Y_1^N$ its output. In a source-coding scenario, $X_1^N$ may be an information source to be compressed and $Y_1^N$ observations available to the decompressor. In either case, there is redundancy in $X_1^N$: added redundancy in channel coding, or removed redundancy in source coding. A good channel code needs to add the least amount of redundancy while still allowing for correct decoding, whereas a good source code eliminates as much redundancy as possible while still allowing reconstruction subject to a distortion criterion.

Polar codes [1] were first developed for binary-input, symmetric, memoryless, channels. They provide a systematic framework to handle this fundamental problem. They are block codes, whose encoding operation consists of an explicit invertible transformation between $X_1^N$ and $U_1^N$. A portion of $U_1^N$ is revealed to the decoder or decompressor. The decoder employs *successive cancellation* (SC) decoding, recovering $U_1^N$ incrementally: first $U_1$, then $U_2$, and so on. Each successive decoding operation uses the observations $Y_1^N$ and the outcome of the previous decoding operations as well as the revealed portion of $U_1^N$. The polarization phenomenon implies that for large enough $N$, the decoding operations polarize to two sets: a 'low entropy' set and a 'high entropy' set. These sets can be determined beforehand, and prescribe which portion of $U_1^N$ to reveal to the decoder or decompressor.

The *rate* of polarization is particularly important for the analysis of polar codes. Their error-free performance at any achievable rate is due to polarization happening sufficiently fast. Fast polarization to the low entropy set for the memoryless setting was established in [1, Theorem 2], [2].

Remarkably, polar codes were extended to a plethora of other memoryless scenarios, including non-binary channels [3], [4], source coding [5], [6], wiretap channels [7], [8], asymmetric channels and sources [9], and more. See the survey paper [10, Section IV] for a large list of extensions and applications. Many of these applications are contingent upon fast polarization to the high-entropy set; for memoryless settings, this was established in [5].

The main tools used for polar code analysis in the memoryless case are the focus of Section III. In particular, we present Arıkan's probabilistic approach, which is at the heart of many polarization results. It is this approach that we extend to settings with memory.

The study of polar codes for scenarios with memory began with [4, Chapter 5]. Şaşoğlu was able to show that polarization indeed occurs for a certain class of processes with memory. In the subsequent work [11] (see also the journal version, [12]), the authors were able to prove polarization for a more general class of processes with memory. One advancement made in that paper was regarding the rate of polarization under memory. The authors showed that polarization to the low entropy set is fast even for processes with memory. Fast polarization to the high entropy set was not addressed.

A practical decoding algorithm for polar codes for FSCs was suggested in [13] (see also [14] for an earlier version, specific to intersymbol interference channels). This algorithm is an extension of SC decoding, taking into account the underlying state structure. Its increase in complexity relative to the complexity of SC decoding is polynomial with the number of states. Thus, it is practical for a moderate number of states. The authors also showed [13, Theorem 3] that their elegant scheme from [9] can be applied to models with memory. To this end, they required the additional assumption of fast polarization both to the low and high entropy sets.

This paper completes the picture. We show that for a large class of processes with memory, polarization is fast both to the low entropy and high entropy sets. Fast polarization to the low entropy set will follow from a specialization of [12]. Fast polarization to the high entropy set, Theorem 13, is the main result of this paper. Consequently, polar codes can be used in settings with memory with vanishing error probability.

This work was supported by the Israel Science Foundation under Grant 1769/13 and by the United States–Israel Binational Science Foundation under grant number 2012016. This work was presented in part in ISIT 2018.

B. Shuval and I. Tal are with the Department of Electrical Engineering, Technion, Haifa 32000, Israel. Email: {bshuval@campus, idotal@ee}.technion.ac.il

Specifically, we consider stationary processes whose memory can be encompassed by an underlying finite-state ergodic[1] Markov chain. This Markov chain governs the joint distribution of $\mathsf{X}_1^N$ and $\mathsf{Y}_1^N$, and is assumed to be hidden. The model is described in detail in Section IV. This family of processes includes, as special cases, finite-state Markov channels [15, Chapter 4.6] with an ergodic state sequence, discrete ergodic sources with finite memory, and many input-constrained systems (e.g., $(d,k)$-runlength limited (RLL) constraint [16], with and without noise).

The tools we develop for this family of processes with memory are the subject of Section V. Our tools mirror those used in the memoryless construction. Thus, we expect that this addition to the 'polar toolbox' will enable natural adaptation of many polar coding results to settings with memory.

## II. Notation

A set of elements is denoted as a list in braces, e.g., $\{1, 2, \ldots, L\}$. The number of elements in a set $A$ is denoted by $|A|$. The disjoint union of two sets $A_0, A_1$ is denoted by $A_0 \uplus A_1$. To use this notation, $A_0$ and $A_1$ must indeed be disjoint. Open and closed intervals are denoted by $(a, b)$ and $[a, b]$, respectively.

We denote $y_j^k = \begin{bmatrix} y_j & y_{j+1} & \cdots & y_k \end{bmatrix}$ for $j < k$. For an arbitrary set of indices $F$ we denote $y_F = \{y_j, j \in F\}$.

In a summation involving multiple variables, if only one variable is being summed, we will make this explicit by underlining it. For example, in $\sum_{\underline{a} \neq b} f(a, b)$ we sum over the values of $a$ that are different than $b$, and $b$ is fixed. In particular, $\sum_{a \neq b} f(a, b) = \sum_b \sum_{\underline{a} \neq b} f(a, b)$.

For a sequence of binary numbers $B_1, B_2, \ldots, B_n$ we define $(B_1 B_2 \cdots B_n)_2 \triangleq \sum_{j=1}^n B_j 2^{n-j}$. Thus, the rightmost digit $B_n$ is the least significant bit. Addition of binary numbers is assumed to be an XOR operation (i.e., modulo-2 addition).

The probability of an event $A$ is denoted by $\mathbb{P}(A)$. Random variables are denoted using a sans-serif font, e.g., $\mathsf{X}$ and their realizations using lower-case letters, e.g., $x$. The distribution of random variable $\mathsf{X}$ is denoted by $P_{\mathsf{X}} = P_{\mathsf{X}}(x)$. When marginalizing distributions, we will sometimes use the shorthand $\sum_x P_{\mathsf{X},\mathsf{Y}} \equiv \sum_x P_{\mathsf{X},\mathsf{Y}}(x, y)$; the summation variable will denote which random variable is being marginalized. The expectation of $\mathsf{X}$ is denoted by $\mathbb{E}[\mathsf{X}]$.

## III. The Polar Toolbox

### A. Various Parameters of Distributions

In this section we introduce several parameters that may be computed from the joint distribution of two random variables: probability of error, Bhattacharyya parameter, conditional entropy, and total variation distance. These parameters are useful for the analysis of polar codes. These parameters are *not* random variables; they are deterministic quantities computed from the joint distribution.

Consider a pair of random variables $(\mathsf{U}, \mathsf{Q})$ with joint distribution $P_{\mathsf{U},\mathsf{Q}}(u, q) = P_{\mathsf{Q}}(q) P_{\mathsf{U}|\mathsf{Q}}(u|q)$. The random variable $\mathsf{U}$ is binary[2] and $\mathsf{Q}$ is some observation dependent on $\mathsf{U}$ that takes values in a finite alphabet $\mathcal{Q}$.

**Definition 1** (Probability of error)**.** The *probability of error* $\mathcal{P}_e(\mathsf{U}|\mathsf{Q})$ of optimally estimating $\mathsf{U}$ from the observation $\mathsf{Q}$, in the sense of minimizing the probability of error, is given by

$$\begin{aligned} \mathcal{P}_e(\mathsf{U}|\mathsf{Q}) &= \sum_q \min\{P_{\mathsf{U},\mathsf{Q}}(0, q), P_{\mathsf{U},\mathsf{Q}}(1, q)\} \\ &= \sum_q P_{\mathsf{Q}}(q) \min\{P_{\mathsf{U}|\mathsf{Q}}(0|q), P_{\mathsf{U}|\mathsf{Q}}(1|q)\}. \end{aligned}$$

**Definition 2** (Bhattacharyya parameter)**.** The *Bhattacharyya parameter* of $\mathsf{U}$ given $\mathsf{Q}$, $\mathcal{Z}(\mathsf{U}|\mathsf{Q})$, is defined as

$$\begin{aligned} \mathcal{Z}(\mathsf{U}|\mathsf{Q}) &= 2 \sum_q \sqrt{P_{\mathsf{U},\mathsf{Q}}(0, q) P_{\mathsf{U},\mathsf{Q}}(1, q)} \\ &= 2 \sum_q P_{\mathsf{Q}}(q) \sqrt{P_{\mathsf{U}|\mathsf{Q}}(0|q) P_{\mathsf{U}|\mathsf{Q}}(1|q)}. \end{aligned} \tag{1}$$

**Definition 3** (Total Variation Distance)**.** The *total variation distance* of $\mathsf{U}$ given $\mathsf{Q}$, $\mathcal{K}(\mathsf{U}|\mathsf{Q})$, is defined as

$$\begin{aligned} \mathcal{K}(\mathsf{U}|\mathsf{Q}) &= \sum_q \left| P_{\mathsf{U},\mathsf{Q}}(0, q) - P_{\mathsf{U},\mathsf{Q}}(1, q) \right| \\ &= \sum_q P_{\mathsf{Q}}(q) \left| P_{\mathsf{U}|\mathsf{Q}}(0|q) - P_{\mathsf{U}|\mathsf{Q}}(1|q) \right|. \end{aligned} \tag{2}$$

The parameters defined above are the definitions for the case where $\mathsf{U}$ is binary. They can be extended to the non-binary case, as described in Appendix B. A final parameter we will use is the conditional entropy. Unlike the other parameters, the conditional entropy is also defined when $\mathsf{U}$ takes values in an arbitrary finite alphabet $\mathcal{U}$, not necessarily binary.

**Definition 4** (Conditional Entropy)**.** The *conditional entropy* of $\mathsf{U}$ given $\mathsf{Q}$, $\mathcal{H}(\mathsf{U}|\mathsf{Q})$, is defined as

$$\begin{aligned} \mathcal{H}(\mathsf{U}|\mathsf{Q}) &= -\sum_q \sum_u P_{\mathsf{U},\mathsf{Q}}(u, q) \log_2 \frac{P_{\mathsf{U},\mathsf{Q}}(u, q)}{\sum_u P_{\mathsf{U},\mathsf{Q}}(u, q)} \\ &= -\sum_q P_{\mathsf{Q}}(q) \sum_u P_{\mathsf{U}|\mathsf{Q}}(u|q) \log_2 P_{\mathsf{U}|\mathsf{Q}}(u|q). \end{aligned} \tag{3}$$

It is easily seen that all four parameters take values in $[0, 1]$ when $\mathsf{U}$ is binary. They are all related, as established in the following lemma.

**Lemma 1.** *The total variation distance, probability of error, conditional entropy, and Bhattacharyya parameter are related by*

$$\mathcal{K}(\mathsf{U}|\mathsf{Q}) = 1 - 2\mathcal{P}_e(\mathsf{U}|\mathsf{Q}) \geq 1 - \mathcal{H}(\mathsf{U}|\mathsf{Q}), \tag{4a}$$

$$\mathcal{Z}(\mathsf{U}|\mathsf{Q})^2 \leq \mathcal{H}(\mathsf{U}|\mathsf{Q}) \leq \mathcal{Z}(\mathsf{U}|\mathsf{Q}), \tag{4b}$$

$$\mathcal{K}(\mathsf{U}|\mathsf{Q}) \leq \sqrt{1 - \mathcal{Z}(\mathsf{U}|\mathsf{Q})^2} \leq \sqrt{1 - \mathcal{H}(\mathsf{U}|\mathsf{Q})^2}. \tag{4c}$$

The proof of Lemma 1 is relegated to Appendix A. We note that the right-most inequality of (4b) was also shown in [6, Proposition 2] and the left-most inequality of (4c) was also shown in [1, Appendix A]; our proof of the latter is more

[1] I.e., aperiodic and irreducible.

[2] This assumption is for the sake of simplicity. See Remark 2 at the end of this subsection for a discussion of the implications of non-binary $\mathsf{U}$.

general. Due to (4a), we shall concentrate in the sequel on $\mathcal{K}(\mathsf{U}|\mathsf{Q})$ rather than $\mathcal{P}_e(\mathsf{U}|\mathsf{Q})$.

In [6], Arıkan used the inequality

$$\mathcal{Z}(\mathsf{U}|\mathsf{Q})^2 \le \mathcal{H}(\mathsf{U}|\mathsf{Q}) \le \log_2(1+\mathcal{Z}(\mathsf{U}|\mathsf{Q})) \tag{5}$$

to show that if the Bhattacharyya parameter approaches 0 or 1 then the conditional entropy approaches 0 or 1 as well and vice versa. An alternative proof of this can be had by (4b). This yields

$$\mathcal{Z}(\mathsf{U}|\mathsf{Q})^2 \le \mathcal{H}(\mathsf{U}|\mathsf{Q}) \le \mathcal{Z}(\mathsf{U}|\mathsf{Q}) \le \sqrt{\mathcal{H}(\mathsf{U}|\mathsf{Q})},$$

which indeed implies that the Bhattacharyya parameter and conditional entropy approach 0 and 1 in tandem. This inequality is tighter than (5); however, as discussed in Appendix B, an advantage of inequality (5) is that it has a natural extension to the case where $\mathsf{U}$ is non-binary.

An additional consequence of Lemma 1 is that (a) if $\mathcal{Z}(\mathsf{U}|\mathsf{Q}) \to 0$ or $\mathcal{H}(\mathsf{U}|\mathsf{Q}) \to 0$ then $\mathcal{K}(\mathsf{U}|\mathsf{Q}) \to 1$ and (b) if $\mathcal{Z}(\mathsf{U}|\mathsf{Q}) \to 1$ or $\mathcal{H}(\mathsf{U}|\mathsf{Q}) \to 1$ then $\mathcal{K}(\mathsf{U}|\mathsf{Q}) \to 0$.

*Remark* 1. By combining (4a) and (4b) we obtain

$$1 - 2\mathcal{P}_e(\mathsf{U}|\mathsf{Q}) \ge 1 - \mathcal{H}(\mathsf{U}|\mathsf{Q}) \ge 1 - \mathcal{Z}(\mathsf{U}|\mathsf{Q}).$$

Rearranging, we obtain the well-known bound, $\mathcal{P}_e(\mathsf{U}|\mathsf{Q}) \le \mathcal{Z}(\mathsf{U}|\mathsf{Q})/2$.

The definitions above naturally extend to the case where instead of $\mathsf{Q}$ there are multiple random variables related to $\mathsf{U}$. For example, consider a triplet of random variables $(\mathsf{U},\mathsf{Q},\mathsf{S})$ with joint distribution $P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(u,q,s)$ such that $\mathsf{U}$ is binary and $\mathsf{Q},\mathsf{S}$ take values in finite alphabets $\mathcal{Q},\mathcal{S}$. We call $\mathsf{S}$ the 'state'. Then,

$$\mathcal{K}(\mathsf{U}|\mathsf{Q},\mathsf{S}) = \sum_{q,s} |P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(0,q,s) - P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(1,q,s)|;$$

the remaining parameters are similarly extended. We say that $\mathcal{K}(\mathsf{U}|\mathsf{Q},\mathsf{S})$ is a *state-informed* (SI) version of $\mathcal{K}(\mathsf{U}|\mathsf{Q})$.

How do the SI parameters compare to their non-SI counterparts? For the entropy, the answer lies in [17, Theorem 2.6.5], the well known property that conditioning reduces entropy. In the following lemma, proved in Appendix A, we consider the other parameters as well.

**Lemma 2.** *Let* $(\mathsf{U},\mathsf{Q},\mathsf{S})$ *be a triplet of random variables with joint distribution* $P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(u,q,s)$. *Then*

$$\mathcal{K}(\mathsf{U}|\mathsf{Q}) \le \mathcal{K}(\mathsf{U}|\mathsf{Q},\mathsf{S}), \tag{6a}$$
$$\mathcal{Z}(\mathsf{U}|\mathsf{Q}) \ge \mathcal{Z}(\mathsf{U}|\mathsf{Q},\mathsf{S}), \tag{6b}$$
$$\mathcal{H}(\mathsf{U}|\mathsf{Q}) \ge \mathcal{H}(\mathsf{U}|\mathsf{Q},\mathsf{S}). \tag{6c}$$

*Remark* 2. In this paper, we assume for simplicity that $\mathsf{U}$ is binary. It is possible to extend our results to the non-binary case. To this end, a suitable extension of the distribution parameters is required. The key properties that need to be preserved are (a) that they be bounded between 0 and 1; (b) that they approach their extreme values in tandem; and (c) that they satisfy Lemma 2. In Appendix B we suggest a suitable extension that satisfies these requirements.

## B. *Polarization*

We review some basics of polarization in this section. The concepts introduced here will be useful in the sequel.

*1) General Definitions:* Consider a strictly stationary process $(\mathsf{X}_j,\mathsf{Y}_j)$, $j = 1,2,\ldots$ with a known joint distribution. We assume that $\mathsf{X}_j$ are binary and $\mathsf{Y}_j \in \mathcal{Y}$, where $\mathcal{Y}$ is a finite alphabet. The random variables $\mathsf{X}_j$ are to be estimated from the observations $\mathsf{Y}_j$. In a channel coding setting, $\mathsf{X}_j$ is the input to a channel and $\mathsf{Y}_j$ its output. In a lossless source coding setting [6], $\mathsf{X}_j$ is a data sequence to be compressed and $\mathsf{Y}_j$ is side information available to the decompressor. In a lossy compression setting [5], the compressor takes a source sequence and distorts it to obtain a sequence $\mathsf{X}_1^N$ that is ultimately recovered by the decompressor.[3]

We denote Arıkan's polarization matrix by $G_N = B_N G_2^{\otimes n}$, where $N = 2^n$, $B_N$ is the $N \times N$ bit-reversal matrix, and $G_2 = \begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix}$. Recall that $G_N^{-1} = G_N$. Following [12], we define

$$\mathsf{U}_1^N = \mathsf{X}_1^N G_N, \tag{7a}$$
$$\mathsf{V}_1^N = \mathsf{X}_{N+1}^{2N} G_N, \tag{7b}$$
$$\mathsf{Q}_i = (\mathsf{U}_1^{i-1}, \mathsf{Y}_1^N), \tag{7c}$$
$$\mathsf{R}_i = (\mathsf{V}_1^{i-1}, \mathsf{Y}_{N+1}^{2N}), \tag{7d}$$

where $i = 1,2,\ldots,N$.

Due to the recursive nature of polar codes, the above equations will be key for passing from a block of length $N$ to a block of length $2N$. Indeed, a block of length $2N$ is given by $\mathsf{F}_1^{2N} = \mathsf{X}_1^{2N} G_{2N}$; using the properties of $G_{2N}$ [1, Section VII] we obtain that $\mathsf{F}_{2i} = \mathsf{U}_i + \mathsf{V}_i$ and $\mathsf{F}_{2i+1} = \mathsf{V}_i$, where $1 \le i \le N$. First, however, let us concentrate on a length-$N$ block. For such a block, equations (7a) and (7c) are pertinent. Although we have described several different communication scenarios, they all share the same succinct description that follows.

A certain subset of indices $F \subset \{1,2,\ldots,N\}$ is preselected according to some rule; the set $F$ dictates the performance of the code. When encoding (compressing), one produces a sequence $\mathsf{U}_1^N$. The relationship between the sequence $\mathsf{U}_1^N$ and the sequence $\mathsf{X}_1^N$ is given by (7a). Then, $\mathsf{U}_F$ is made available to the decoder.[4] The decoding (decompressing) operation is iterative. For $i = 1,2,\ldots$, the decoder estimates $\mathsf{U}_i$ from $\mathsf{Q}_i$; it uses its previous estimates of $\mathsf{U}_1^{i-1}$ to form $\mathsf{Q}_i$. Whenever it encounters an index in $F$, it returns as its estimate the relevant value from $\mathsf{U}_F$. After estimating $\mathsf{U}_1^N$, the decoder recovers $\mathsf{X}_1^N$ via (7a).

The polarization phenomenon is that for large enough $n$, the fraction of indices with moderate conditional entropy, $|\{i : \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i) \in (\epsilon, 1-\epsilon)\}|/N$, becomes negligibly small for any $\epsilon > 0$. One approach [1], [6] to derive such results is probabilistic. Rather than counting the number of indices with moderate conditional entropy, a sequence of random variables $\mathsf{H}_n$, $n = 1,2,\ldots$ is defined. The random variable $\mathsf{H}_n$ assumes the

[3] In fact, in a lossy compression setting, with side information known to both compressor and decompressor, the process is $(\mathsf{X}_j,\mathsf{Y}_j)$, where $\mathsf{Y}_j = (\mathsf{Y}'_j,\mathsf{Y}''_j)$. The random variables $\mathsf{Y}'$ are the sequence to be compressed and the random variables $\mathsf{Y}''$ are the side information.

[4] Depending on the application, this can be done either explicitly, by shared randomness, or both.

value $\mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i)$, with $i$ selected uniformly from $\{1, 2, \dots, N\}$. Thus, the probability that $\mathsf{H}_n$ lies in a certain range equals the fraction of indices whose conditional entropies lie in this range.

The recursive nature of the polarization transform is at the heart of the probabilistic approach. Concretely, let $\mathsf{B}_1, \mathsf{B}_2, \dots$ be a sequence of independent and identically distributed (i.i.d.) Bernoulli-$1/2$ random variables. We set $i - 1 = (\mathsf{B}_1\mathsf{B}_2 \cdots \mathsf{B}_n)_2$; indeed, $i$ assumes any value in $\{1, 2, \dots, N\}$ with equal probability. Define the random variables

$$\begin{aligned} \mathsf{K}_n &= \mathcal{K}(\mathsf{U}_i|\mathsf{U}_1^{i-1}, \mathsf{Y}_1^N) = \mathcal{K}(\mathsf{U}_i|\mathsf{Q}_i), \\ \mathsf{Z}_n &= \mathcal{Z}(\mathsf{U}_i|\mathsf{U}_1^{i-1}, \mathsf{Y}_1^N) = \mathcal{Z}(\mathsf{U}_i|\mathsf{Q}_i), \\ \mathsf{H}_n &= \mathcal{H}(\mathsf{U}_i|\mathsf{U}_1^{i-1}, \mathsf{Y}_1^N) = \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i) \end{aligned} \tag{8}$$

whenever $(i - 1) = (\mathsf{B}_1\mathsf{B}_2 \cdots \mathsf{B}_n)_2$. That is, they denote the relevant distribution parameters for a uniformly chosen index after $n$ polarization steps. We call $\mathsf{K}_n, \mathsf{Z}_n$, and $\mathsf{H}_n$, $n = 1, 2, \dots$ the *total variation distance process*, the *Bhattacharyya process*, and the *conditional entropy process*, respectively.

When passing from a length-$N$ block to a block of length $2N$, by the properties of $G_N$ [1, Section VII],

$$\mathsf{K}_{n+1} = \begin{cases} \mathcal{K}(\mathsf{U}_i + \mathsf{V}_i|\mathsf{Q}_i, \mathsf{R}_i), & \text{if } \mathsf{B}_{n+1} = 0, \\ \mathcal{K}(\mathsf{V}_i|\mathsf{U}_i + \mathsf{V}_i, \mathsf{Q}_i, \mathsf{R}_i), & \text{if } \mathsf{B}_{n+1} = 1. \end{cases} \tag{9}$$

Similar relationships hold for $\mathsf{H}_{n+1}$ and $\mathsf{Z}_{n+1}$. We shall use the mnemonics $\mathsf{K}_n^-$ and $\mathsf{K}_n^+$ to denote $\mathcal{K}(\mathsf{U}_i + \mathsf{V}_i|\mathsf{Q}_i, \mathsf{R}_i)$ and $\mathcal{K}(\mathsf{V}_i|\mathsf{U}_i + \mathsf{V}_i, \mathsf{Q}_i, \mathsf{R}_i)$, respectively. I.e., $\mathsf{K}_{n+1}$ assumes the value $\mathsf{K}_n^-$ when $\mathsf{B}_{n+1} = 0$ and the value $\mathsf{K}_n^+$ when $\mathsf{B}_{n+1} = 1$. We shall use similar mnemonics for $\mathsf{H}_n$ and $\mathsf{Z}_n$.

The probability law of $(\mathsf{U}_i, \mathsf{V}_i, \mathsf{Q}_i, \mathsf{R}_i)$ can be obtained from the probability law of $(\mathsf{X}_1^{2N}, \mathsf{Y}_1^{2N})$ using (7). Moreover, for fixed $i$, there exists a function $f$, which depends solely on $i$, such that

$$\begin{aligned} (\mathsf{U}_i, \mathsf{Q}_i) &= f(\mathsf{X}_1^N, \mathsf{Y}_1^N), \\ (\mathsf{V}_i, \mathsf{R}_i) &= f(\mathsf{X}_{N+1}^{2N}, \mathsf{Y}_{N+1}^{2N}). \end{aligned} \tag{10}$$

This can be seen by comparing (7a) and (7c) with (7b) and (7d). Due to stationarity, $P_{\mathsf{U}_i,\mathsf{Q}_i} = P_{\mathsf{V}_i,\mathsf{R}_i}$.

Denote $\mathsf{T}_i = \mathsf{U}_i + \mathsf{V}_i$, as in Figure 1. The mapping $(\mathsf{U}_i, \mathsf{V}_i) \mapsto (\mathsf{T}_i, \mathsf{V}_i)$ is one-to-one and onto. Hence,

$$P_{\mathsf{T}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i}(t, v, q, r) = P_{\mathsf{U}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i}(t + v, v, q, r). \tag{11}$$

We now formally define polarization and fast polarization.

**Definition 5.** Let $\mathsf{A}_n$, $n = 1, 2, \dots$ be a sequence of random variables that take values in $[0, 1]$.

1) The sequence $\mathsf{A}_n$ *polarizes* if it converges almost surely to a $\{0, 1\}$-random variable $\mathsf{A}_\infty$ as $n \to \infty$. We will sometimes abbreviate this by saying that "$\mathsf{A}_n$ polarizes to $\mathsf{A}_\infty$."
2) The sequence $\mathsf{A}_n$ *polarizes fast to* 0 with $\beta > 0$ if it polarizes and
$$\lim_{n\to\infty} \mathbb{P}\left(\mathsf{A}_n < 2^{-2^{n\beta}}\right) = \mathbb{P}\left(\mathsf{A}_\infty = 0\right).$$
3) The sequence $\mathsf{A}_n$ *polarizes fast to* 1 with $\beta > 0$ if it polarizes and
$$\lim_{n\to\infty} \mathbb{P}\left(\mathsf{A}_n > 1 - 2^{-2^{n\beta}}\right) = \mathbb{P}\left(\mathsf{A}_\infty = 1\right).$$

When the precise value of $\beta$ is either obvious from the context or not needed, we will write that $\mathsf{A}_n$ polarizes fast to, say, 0, without mentioning the value of $\beta$.

The following lemma, first obtained by Arıkan and Telatar in [2] and later adapted to the general case by Şaşoğlu in [4], is an important tool for establishing fast polarization for a sequence of random variables.

**Lemma 3.** *[2],[4, Lemma 4.2] Let $\mathsf{B}_n$, $n = 1, 2, \dots$ be an i.i.d. Bernoulli-$1/2$ process and $\mathsf{A}_n$, $n = 1, 2, \dots$ be a $[0, 1]$-valued process that polarizes to a $\{0, 1\}$-random variable $\mathsf{A}_\infty$. Assume that there exist $k \geq 1$ and $d_0, d_1 > 0$ such that for $i = 0, 1$,*

$$\mathsf{A}_{n+1} \leq k\mathsf{A}_n^{d_i} \quad \text{if } \mathsf{B}_{n+1} = i.$$

*Then, for any $0 < \beta < E = (\log_2 d_0 + \log_2 d_1)/2$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathsf{A}_n < 2^{-2^{n\beta}}\right) = \mathbb{P}\left(\mathsf{A}_\infty = 0\right). \tag{12}$$

*Remark* 3. It was shown in [18] that Lemma 3 can be strengthened. Namely, equation (12) can be replaced with the stronger assertion $\lim_{n_0\to\infty} \mathbb{P}(\mathsf{A}_n \leq 2^{-2^{n\beta}} \text{ for all } n \geq n_0) = \mathbb{P}(\mathsf{A}_\infty = 0)$. Hence, any result based on Lemma 3, such as Theorems 7 and 13, can be strengthened similarly.

*2) The Memoryless Case:* The memoryless case is characterized by $P_{\mathsf{X}_1^N,\mathsf{Y}_1^N}(x_1^N, y_1^N) = \prod_{j=1}^N P_{\mathsf{X},\mathsf{Y}}(x_j, y_j)$. Arıkan showed in [1] that in the memoryless case the process $\mathsf{H}_n$ polarizes. Consequently, when $n$ is large enough, for all but a negligible fraction of indices $i$, $\mathcal{H}(\mathsf{U}_i|\mathsf{U}_1^{i-1}, \mathsf{Y}_1^N)$ is either very close to 0 or very close to 1.

To achieve this, Arıkan had shown that the sequence $\mathsf{H}_n$, $n = 1, 2, \dots$ is a bounded martingale sequence and thus converges almost surely to some random variable $\mathsf{H}_\infty$. By showing that $\mathsf{H}_\infty$ can only assume the values 0 and 1, polarization is obtained.

The Bhattacharyya process, in the memoryless case, is a bounded supermartingale that converges almost surely to a $\{0, 1\}$-random variable $\mathsf{Z}_\infty$. The process $\mathsf{Z}_n$ satisfies Lemma 3 with $E = 1/2$ by virtue of [1, Proposition 5], by which

$$\mathsf{Z}_{n+1} = \begin{cases} \leq 2\mathsf{Z}_n, & \text{if } \mathsf{B}_{n+1} = 0, \\ \mathsf{Z}_n^2, & \text{if } \mathsf{B}_{n+1} = 1. \end{cases}$$

Thus, the Bhattacharyya process polarizes fast to 0 with any $\beta < 1/2$.

Fast polarization of the Bhattacharyya parameter is important for the performance analysis of polar codes. In particular, this was instrumental in Arıkan's proof that polar codes are capacity-achieving for binary-input, memoryless, symmetric, channels [1]. Arıkan had upper-bounded the probability of error of polar codes by the union-Bhattacharyya bound. Thanks to fast polarization of the Bhattacharyya process to 0, the bound converges to 0.

The additional requirement of fast polarization of $\mathsf{Z}_n$ to 1 is important for many applications of polar codes. For example, it is integral to source coding applications [5] and to channel coding without symmetry assumptions [9]. In [5, Theorem 16], this fast polarization was established by showing that the process $\tilde{\mathsf{Z}}_n = 1 - \mathsf{Z}_n^2$ polarizes fast to 0 with $\beta < 1/2$. Another way to see this, which we pursue in the sequel, is via the total variation process $\mathsf{K}_n$.

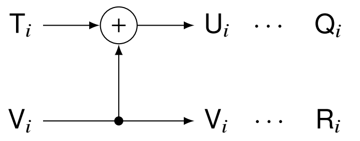
$\mathsf{T}_i$
+
$\mathsf{U}_i$ ··· $\mathsf{Q}_i$
$\mathsf{V}_i$
$\mathsf{V}_i$ ··· $\mathsf{R}_i$

Fig. 1. Illustration of a polarization transform. Random variables $(\mathsf{U}_i, \mathsf{Q}_i)$ have joint distribution $P_{\mathsf{U}_i,\mathsf{Q}_i}$ and random variables $(\mathsf{V}_i, \mathsf{R}_i)$ have joint distribution $P_{\mathsf{V}_i,\mathsf{R}_i}$.

A consequence of Lemma 1 is that if $\mathsf{K}_n$ polarizes fast to 0 then $\mathsf{Z}_n$ must polarize fast to 1. The total variation process $\mathsf{K}_n$ can be shown to polarize (we show this in Corollary 11 for a more general setting). Fast polarization of $\mathsf{K}_n$ to 0 is obtained from Lemma 3 and the following proposition.

**Proposition 4.** *Assume that* $(\mathsf{X}_j, \mathsf{Y}_j)$, $j \in \mathbb{Z}$ *is a memoryless process, where* $\mathsf{X}_j$ *is binary and* $\mathsf{Y}_j \in \mathcal{Y}$. *Then,*

$$\mathsf{K}_{n+1} = \begin{cases} \mathsf{K}_n^2, & \text{if } \mathsf{B}_{n+1} = 0, \\ \leq 2\mathsf{K}_n, & \text{if } \mathsf{B}_{n+1} = 1. \end{cases} \tag{13}$$

In the sequel, we shall generalize this proposition to a non-memoryless case. The proof for the memoryless case serves as preparation for the more general case, which uses similar techniques. For an extension of Proposition 4 to the case where $\mathsf{X}_j$ is non-binary, see Appendix B.

*Proof:* Fix $\mathsf{B}_1, \ldots, \mathsf{B}_n$ and let $i - 1 = (\mathsf{B}_1\mathsf{B}_2 \cdots \mathsf{B}_n)_2$. This also fixes the value of $\mathsf{K}_n$. Using (10) and the memoryless assumption, we denote $P \equiv P_{\mathsf{U}_i,\mathsf{Q}_i} = P_{\mathsf{V}_i,\mathsf{R}_i}$, by which

$$P_{\mathsf{U}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i}(u, v, q, r) = P(u, q)P(v, r).$$

Note that $\mathsf{K}_n = \mathcal{K}(\mathsf{U}_i|\mathsf{Q}_i) = \mathcal{K}(\mathsf{V}_i|\mathsf{R}_i)$.

Set $\mathsf{T}_i = \mathsf{U}_i + \mathsf{V}_i$; by (11),

$$P_{\mathsf{T}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i}(t, v, q, r) = P(t + v, q)P(v, r),$$

and $P_{\mathsf{T}_i,\mathsf{Q}_i,\mathsf{R}_i}(t, q, r) = \sum_{v=0}^{1} P_{\mathsf{T}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i}(t, v, q, r)$. A single-step polarization from $\mathsf{K}_n$ to $\mathsf{K}_{n+1}$, (9), becomes

$$\mathsf{K}_{n+1} = \begin{cases} \mathcal{K}(\mathsf{T}_i|\mathsf{Q}_i, \mathsf{R}_i), & \text{if } \mathsf{B}_{n+1} = 0, \\ \mathcal{K}(\mathsf{V}_i|\mathsf{T}_i, \mathsf{Q}_i, \mathsf{R}_i), & \text{if } \mathsf{B}_{n+1} = 1. \end{cases} \tag{14}$$

Assume first that $\mathsf{B}_{n+1} = 0$. Then

$$\begin{aligned} \mathsf{K}_{n+1} &= \sum_{q,r} \left| P_{\mathsf{T}_i,\mathsf{Q}_i,\mathsf{R}_i}(0, q, r) - P_{\mathsf{T}_i,\mathsf{Q}_i,\mathsf{R}_i}(1, q, r) \right| \\ &= \sum_{q,r} \left| \sum_{v=0}^{1} P(v, r)(P(v, q) - P(v + 1, q)) \right| \\ &= \sum_{q,r} \left| \Big(P(0, q) - P(1, q)\Big)\Big(P(0, r) - P(1, r)\Big) \right| \\ &\overset{\text{(a)}}{=} \sum_{q,r} |P(0, q) - P(1, q)| \cdot |P(0, r) - P(1, r)| \\ &= \sum_{q} |P(0, q) - P(1, q)| \cdot \sum_{r} |P(0, r) - P(1, r)| \\ &= \mathsf{K}_n^2, \end{aligned}$$

where (a) is because $|ab| = |a| \cdot |b|$ for any two numbers $a, b$. Next, assume that $\mathsf{B}_{n+1} = 1$. Observe that for any four numbers $a, b, c, d$,

$$(ab - cd) = \frac{(a + c)(b - d) + (b + d)(a - c)}{2}. \tag{15}$$

With a slight abuse of notation, we denote $P(q) = P_{\mathsf{Q}_i}(q) = P(0, q) + P(1, q)$. Then, $P(r) = P_{\mathsf{R}_i}(r) = P(0, r) + P(1, r)$. Thus,

$$\begin{aligned} \mathsf{K}_{n+1} &= \sum_{t,q,r} \left| P_{\mathsf{T}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i}(t, 0, q, r) - P_{\mathsf{T}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i}(t, 1, q, r) \right| \\ &= \sum_{t,q,r} |P(t, q)P(0, r) - P(t + 1, q)P(1, r)| \\ &\leq \frac{1}{2} \sum_{t,q,r} P(q) \, |P(0, r) - P(1, r)| \\ &\quad + \frac{1}{2} \sum_{t,q,r} P(r) \, |P(t, q) - P(t + 1, q)| \\ &= \frac{1}{2} \sum_{t,r} |P(0, r) - P(1, r)| \\ &\quad + \frac{1}{2} \sum_{t,q} |P(t, q) - P(t + 1, q)| \\ &= 2\mathsf{K}_n, \end{aligned}$$

where the inequality is due to a combination of (15) with the triangle inequality.

We have shown that $\mathsf{K}_{n+1} = \mathsf{K}_n^2$ if $\mathsf{B}_{n+1} = 0$ and $\mathsf{K}_{n+1} \leq 2\mathsf{K}_n$ if $\mathsf{B}_{n+1} = 1$, completing the proof. ■

*Remark* 4. Several other authors have independently looked at the polarization of the total variation distance. For example, [19, Proposition 5.1] derives relations similar to (13); the top equality of (13) is also shown in [20, Equation 12]. Those results were derived for binary-input, memoryless, and symmetric channels. Our Proposition 4, on the other hand, does not require symmetry. We note in passing that it is also easily extended to a non-stationary case (similar to [21, Appendix 2.A] for the Bhattacharyya process), but that is outside the scope of this paper.

## IV. Finite-State Aperiodic Irreducible Markov Processes

In this section we introduce a class of processes with memory that we call *Finite-state Aperiodic Irrecducible Markov processes* (FAIM processes). This is the class of processes for which we establish polarization and fast polarization.

These processes are described using an underlying state sequence. Often, however, the state sequence is hidden. The polarization results we obtain apply to processes with a hidden state sequence.

### A. Definition

Let $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$, $j \in \mathbb{Z}$ be a strictly stationary process, where $\mathsf{X}_j$ is binary, $\mathsf{Y}_j \in \mathcal{Y}$, and $\mathsf{S}_j \in \mathcal{S}$. The alphabets $\mathcal{Y}$ and $\mathcal{S}$ are finite; in particular, $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$. We call $\mathsf{S}_j, j \in \mathbb{Z}$ the *state sequence*; it governs the distribution of sequences $\mathsf{X}_j$ and $\mathsf{Y}_j$, $j \in \mathbb{Z}$.

We may think of $\mathsf{X}_j$ as a state-dependent input to a state-dependent channel with output $\mathsf{Y}_j$. Alternatively, $\mathsf{X}_j$ may be some state-dependent source to be compressed, and $\mathsf{Y}_j$ an observation that the decoder may use as a decompression aid. The state sequence encompasses the memory of the process.

The process is described by the conditional probability $P_{\mathsf{X}_j,\mathsf{Y}_j,\mathsf{S}_j|\mathsf{S}_{j-1}}$, which, by the stationarity assumption, is independent of $j$. We assume a Markov property: conditioned on $\mathsf{S}_{j-1}$, the random variables $\mathsf{X}_k, \mathsf{Y}_k, \mathsf{S}_k$ are independent of $\mathsf{X}_l, \mathsf{Y}_l, \mathsf{S}_{l-1}$ for any $l < j \leq k$. Thus, for any $N > M > 0$,

$$
\begin{aligned}
&P_{\mathsf{X}_1^N,\mathsf{Y}_1^N,\mathsf{S}_N|\mathsf{S}_0} \\
&\quad= \sum_b P_{\mathsf{X}_1^M,\mathsf{Y}_1^M,\mathsf{S}_M,\mathsf{X}_{M+1}^N,\mathsf{Y}_{M+1}^N,\mathsf{S}_N|\mathsf{S}_0} \\
&\quad= \sum_b P_{\mathsf{X}_{M+1}^N,\mathsf{Y}_{M+1}^N,\mathsf{S}_N|\mathsf{S}_M,\mathsf{X}_1^M,\mathsf{Y}_1^M,\mathsf{S}_0} \cdot P_{\mathsf{X}_1^M,\mathsf{Y}_1^M,\mathsf{S}_M|\mathsf{S}_0} \\
&\quad= \sum_b P_{\mathsf{X}_{M+1}^N,\mathsf{Y}_{M+1}^N,\mathsf{S}_N|\mathsf{S}_M} \cdot P_{\mathsf{X}_1^M,\mathsf{Y}_1^M,\mathsf{S}_M|\mathsf{S}_0},
\end{aligned}
\tag{16}
$$

where $b$ in the sum represents the value of the middle state $\mathsf{S}_M$.

The state sequence is a finite-state homogeneous Markov chain. We denote its marginal distribution by $\pi$, and use the shorthand

$$
\begin{aligned}
\pi_N(a) &= P_{\mathsf{S}_N}(a) \\
\pi_{N|M}(b|a) &= P_{\mathsf{S}_N|\mathsf{S}_M}(b|a) \\
\pi_{N,M}(b,a) &= P_{\mathsf{S}_N,\mathsf{S}_M}(b,a),
\end{aligned}
\tag{17}
$$

where $N > M$. Note that $\pi_N(a) = \pi_0(a)$ and $\pi_{N|M}(b|a) = \pi_{N-M|0}(b|a)$.

A finite-state homogeneous Markov chain is aperiodic and irreducible (ergodic) if and only if there is some $N_0 > 0$ such that for any $N \geq N_0$, $\pi_{N|0}(b|a) > 0$ for any $a, b \in \mathcal{S}$. It can be shown that it has a unique stationary distribution $\pi_0$ and $\pi_0(a) > 0$ for any $a \in \mathcal{S}$. Moreover, $\pi_{N|0}(b|a) \to \pi_0(b)$ exponentially fast as $N \to \infty$ for any $a, b \in \mathcal{S}$. See, e.g., [22, Section 8].

The process $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$, $j \in \mathbb{Z}$ is called a *finite-state aperiodic irreducible Markov* process if the underlying Markov process $\mathsf{S}_j$, $j \in \mathbb{Z}$ is homogenous, finite-state, strictly stationary, aperiodic, and irreducible.[5] In the sequel, we assume that $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$, $j \in \mathbb{Z}$ is a FAIM process.

At this point, the reader may wonder why we have imposed aperiodicity and irreducibility. In [12, Theorem 4], it was demonstrated that periodic processes may not polarize. We assume aperiodicity to ensure that polarization indeed happens. As for irreducibility, note that since the number of states is finite, the state sequence $\mathsf{S}_j$, $j \in \mathbb{Z}$ must reach an irreducible sink after sufficient time. Hence, the irreducibility assumption is equivalent to assuming that the state sequence begins in some irreducible sink.

Our model applies to many problems in information theory that can be described using states. For example, compression of finite memory sources and coding for input constrained channels. Additionally, our model may be applied to finite-state channels; in this case, the FAIM state sequence describes both the channel state and input state. That is, FAIM processes enable us to model non-i.i.d. input sequences.

One famous example of a finite state model is the indecomposable FSC model considered in [15, Section 4.6]. There are some differences between this model and ours. Most importantly, a FAIM process has a specified input distribution, whereas an indecomposable FSC is devoid of such specification. Instead, an indecomposable FSC imposes conditions that should hold for all input sequences. That said, once a hidden Markov input distribution has been specified, we can define a process in which the state space is the Cartesian product of the state spaces of the input distribution and the channel. In many important cases, e.g. a Gilbert-Elliot channel [23], this combined process falls under the FAIM framework.

### *B. Blocks of a FAIM Process*

Typically, the state sequence is not observed. The joint distribution of $(\mathsf{X}_1^N, \mathsf{Y}_1^N)$ is given by

$$
P_{\mathsf{X}_1^N,\mathsf{Y}_1^N}(x_1^N, y_1^N) = \sum_{b,a} P_{\mathsf{X}_1^N,\mathsf{Y}_1^N,\mathsf{S}_N|\mathsf{S}_0}(x_1^N, y_1^N, b|a)\pi_0(a),
$$

where $\pi_0$ is the stationary distribution of the initial state.

**Definition 6** (Block)**.** Let $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$, $j \in \mathbb{Z}$ be a FAIM process and assume $M > L$. We call $(\mathsf{X}_{L+1}^M, \mathsf{Y}_{L+1}^M)$ a *block* of the FAIM process. Its length is $M - L$.

State $\mathsf{S}_L$ is called the *initial* state of the block. State $\mathsf{S}_M$ is called the *final* state of the block.

We emphasize that the initial state of the block $(\mathsf{X}_{L+1}^M, \mathsf{Y}_{L+1}^M)$ is $\mathsf{S}_L$ and *not* $\mathsf{S}_{L+1}$.

The following lemma holds for any two non-overlapping blocks of a FAIM process. It establishes that FAIM processes are a special case of the family of processes considered in [12].

**Lemma 5.** *Assume that* $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$*,* $j \in \mathbb{Z}$ *is a FAIM process. Then, there exists a non-increasing sequence* $\psi(N)$*,* $\psi(N) \to 1$ *as* $N \to \infty$*, such that for any* $N > M \geq L \geq 1$*,*

$$
P_{\mathsf{X}_1^L,\mathsf{Y}_1^L,\mathsf{X}_{M+1}^N,\mathsf{Y}_{M+1}^N} \leq \psi(M-L) \cdot P_{\mathsf{X}_1^L,\mathsf{Y}_1^L} \cdot P_{\mathsf{X}_{M+1}^N,\mathsf{Y}_{M+1}^N}, \tag{18}
$$

*and* $\psi(0) < \infty$*.*

We relegate the proof to Appedix C. We remark, however, that

$$
\psi(N) = \begin{cases} \max\limits_{a,b} \dfrac{\pi_{N|0}(b|a)}{\pi_0(b)}, & \text{if } N > 0, \\ \max\limits_{a} \dfrac{1}{\pi_0(a)}, & \text{if } N = 0. \end{cases} \tag{19}
$$

I.e., $\psi(\cdot)$ is completely determined by the distribution of the underlying state sequence. Indeed, $\psi(N) \to 1$ as $N \to \infty$.

A process satisfying (18) with $\psi(N) \to 1$ as $N \to \infty$ is called *$\psi$-mixing*.[6] The function $\psi(\cdot)$ is called the *mixing coefficient*. The operational meaning of (18) is that as $L$ and $M$ becomes

[5] We remark that the process $(\mathsf{X}_j, \mathsf{Y}_j)$, $j \in \mathbb{Z}$ is not necessarily Markov.

[6] In some literature, e.g. [24], the term used is $\psi^*$-mixing.

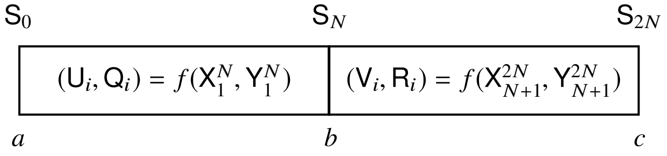
$\mathsf{S}_0$
$\mathsf{S}_N$
$\mathsf{S}_{2N}$
$(\mathsf{U}_i, \mathsf{Q}_i) = f(\mathsf{X}_1^N, \mathsf{Y}_1^N)$
$(\mathsf{V}_i, \mathsf{R}_i) = f(\mathsf{X}_{N+1}^{2N}, \mathsf{Y}_{N+1}^{2N})$
$a$
$b$
$c$

Fig. 2. Two adjacent length-$N$ blocks of a FAIM process. When $i - 1 = (\mathsf{B}_1\mathsf{B}_2 \cdots \mathsf{B}_n)_2$, there is a function $f$ such that $(\mathsf{U}_i, \mathsf{Q}_i) = f(\mathsf{X}_1^N, \mathsf{Y}_1^N)$ and $(\mathsf{V}_i, \mathsf{R}_i) = f(\mathsf{X}_{N+1}^{2N}, \mathsf{Y}_{N+1}^{2N})$. The initial state of the first block, $\mathsf{S}_0$, assumes value $a \in \mathcal{S}$. The final state of the first block, $\mathsf{S}_N$, which is also the initial state of the second block, assumes value $b \in \mathcal{S}$. The final state of the second block, $\mathsf{S}_{2N}$, assumes value $c \in \mathcal{S}$.

more separated in time, the blocks $(\mathsf{X}_1^L, \mathsf{Y}_1^L)$ and $(\mathsf{X}_{M+1}^N, \mathsf{Y}_{M+1}^N)$ become almost independent.[7]

Two adjacent blocks of the process share a state. The final state of the first block is the initial state of the second block. Given the shared state, the two blocks are independent. We capture this in the following lemma.

**Lemma 6.** *For any $N > M \geq 1$,*

$$P_{\mathsf{X}_1^M,\mathsf{Y}_1^M,\mathsf{X}_{M+1}^N,\mathsf{Y}_{M+1}^N|\mathsf{S}_M} = P_{\mathsf{X}_1^M,\mathsf{Y}_1^M|\mathsf{S}_M} \cdot P_{\mathsf{X}_{M+1}^N,\mathsf{Y}_{M+1}^N|\mathsf{S}_M}, \tag{20a}$$

$$P_{\mathsf{X}_1^M,\mathsf{Y}_1^M,\mathsf{X}_{M+1}^N,\mathsf{Y}_{M+1}^N|\mathsf{S}_0,\mathsf{S}_M,\mathsf{S}_N} = P_{\mathsf{X}_1^M,\mathsf{Y}_1^M|\mathsf{S}_0,\mathsf{S}_M} \cdot P_{\mathsf{X}_{M+1}^N,\mathsf{Y}_{M+1}^N|\mathsf{S}_M,\mathsf{S}_N}. \tag{20b}$$

This is a direct consequence of the Markov property. A formal derivation can be found in Appendix C.

A notational convention concludes this section. Our analysis involves the use of some states of blocks of a FAIM process. We will use ascending letters to denote values of ordered states. That is, a state with value $a$ occurs before a state with value $b$, which, in turn, occurs before a state with value $c$. In Figure 2 we illustrate a particular case that will be used in the sequel. A block of length $2N$ comprises two adjacent blocks of length $N$. State $\mathsf{S}_0$, the initial state of the first block, may take value $a$, state $\mathsf{S}_N$, at the end of the first block and the beginning of the second block, may take value $b$, and state $\mathsf{S}_{2N}$, at the end of the second block, may take value $c$. We emphasize that $a, b, c \in \mathcal{S}$ are *not* random variables, but *values* of the relevant states.

### C. Boundary-State-Informed Parameters for FAIM Processes

Let $(\mathsf{X}_1^N, \mathsf{Y}_1^N)$ be a block of a FAIM process with state sequence $\mathsf{S}_j$. Let $f(\cdot,\cdot)$ be some function independent of the state sequence such that

$$(\mathsf{U}, \mathsf{Q}) = f(\mathsf{X}_1^N, \mathsf{Y}_1^N)$$

and $\mathsf{U}$ is binary. We denote

$$P_a^b(u, q) \triangleq P_{\mathsf{U},\mathsf{Q}|\mathsf{S}_N,\mathsf{S}_0}(u, q|b, a) = \frac{P_{\mathsf{U},\mathsf{Q},\mathsf{S}_N|\mathsf{S}_0}(u, q, b|a)}{\pi_{N|0}(b|a)}. \tag{21}$$

I.e., this is the distribution of $\mathsf{U}$ and $\mathsf{Q}$, functions of a block of length $N$, conditioned on the initial state being $\mathsf{S}_0 = a$ and the final state being $\mathsf{S}_N = b$. We further define

$$P_a^b(q) = P_a^b(0, q) + P_a^b(1, q). \tag{22}$$

I.e., $P_a^b(q) = P_{\mathsf{Q}|\mathsf{S}_N,\mathsf{S}_0}(q|b, a)$, and $\sum_q P_a^b(q) = 1$.

We denote the results of replacing $P_{\mathsf{U},\mathsf{Q}}(u, q)$ with $P_a^b(u, q)$ in Equations (1) to (3) by $\mathcal{Z}_a^b(\mathsf{U}|\mathsf{Q})$, $\mathcal{K}_a^b(\mathsf{U}|\mathsf{Q})$, and $\mathcal{H}_a^b(\mathsf{U}|\mathsf{Q})$, respectively. For example,

$$\mathcal{K}_a^b(\mathsf{U}|\mathsf{Q}) = \sum_q \left|P_a^b(0, q) - P_a^b(1, q)\right|. \tag{23}$$

Since $P_{\mathsf{U},\mathsf{Q},\mathsf{S}_N,\mathsf{S}_0}(u, q, b, a) = P_a^b(u, q) \cdot \pi_{N,0}(b, a)$, we have $\mathcal{K}(\mathsf{U}|\mathsf{Q}, \mathsf{S}_N, \mathsf{S}_0) = \sum_{a,b} \pi_{N,0}(b, a)\mathcal{K}_a^b(\mathsf{U}|\mathsf{Q})$. This leads to the following definition.

**Definition 7.** Let $(\mathsf{U}, \mathsf{Q}) = f(\mathsf{X}_1^N, \mathsf{Y}_1^N)$ with $\mathsf{U}$ binary. The *boundary-state-informed* (BSI) total variation distance, Bhattacharyya parameter, and conditional entropy are respectively defined as

$$\mathcal{K}(\mathsf{U}|\mathsf{Q}, \mathsf{S}_N, \mathsf{S}_0) = \sum_{a,b} \pi_{N,0}(b, a)\mathcal{K}_a^b(\mathsf{U}|\mathsf{Q}),$$
$$\mathcal{Z}(\mathsf{U}|\mathsf{Q}, \mathsf{S}_N, \mathsf{S}_0) = \sum_{a,b} \pi_{N,0}(b, a)\mathcal{Z}_a^b(\mathsf{U}|\mathsf{Q}),$$
$$\mathcal{H}(\mathsf{U}|\mathsf{Q}, \mathsf{S}_N, \mathsf{S}_0) = \sum_{a,b} \pi_{N,0}(b, a)\mathcal{H}_a^b(\mathsf{U}|\mathsf{Q}).$$

BSI parameters are defined for blocks of the process; they depend on the initial and final states of the block. Invoking (6) we relate the distribution parameters to their BSI counterparts,

$$\begin{aligned} \mathcal{K}(\mathsf{U}|\mathsf{Q}) &\leq \mathcal{K}(\mathsf{U}|\mathsf{Q}, \mathsf{S}_N, \mathsf{S}_0), \\ \mathcal{Z}(\mathsf{U}|\mathsf{Q}) &\geq \mathcal{Z}(\mathsf{U}|\mathsf{Q}, \mathsf{S}_N, \mathsf{S}_0), \\ \mathcal{H}(\mathsf{U}|\mathsf{Q}) &\geq \mathcal{H}(\mathsf{U}|\mathsf{Q}, \mathsf{S}_N, \mathsf{S}_0). \end{aligned} \tag{24}$$

## V. Fast Polarization for FAIM Processes

This section contains our main result: fast polarization for FAIM processes. First, we show that they polarize by leveraging the results of [12]. Then, we show fast polarization of the Bhattacharyya parameter and of the total variation distance to zero.

The notation of Section III-B holds, without change, for FAIM processes. That is, $\mathsf{U}_1^N, \mathsf{V}_1^N, \mathsf{Q}_i, \mathsf{R}_i$, $i = 1, \ldots, N$ are defined using (7). The random variables $\mathsf{B}_1, \ldots, \mathsf{B}_n$ are used for a random, iterative, uniform selection of an index after $n$ polarization steps. That is, they constitute the binary expansion of $i - 1$, through which the random variables $\mathsf{K}_n = \mathcal{K}(\mathsf{U}_i|\mathsf{Q}_i)$, $\mathsf{H}_n = \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i)$, and $\mathsf{Z}_n = \mathcal{Z}(\mathsf{U}_i|\mathsf{Q}_i)$ are defined. Random variable $\mathsf{K}_{n+1}$ is related to $\mathsf{K}_n$ by (9). I.e., $\mathsf{K}_{n+1} = \mathsf{K}_n^-$ if $\mathsf{B}_{n+1} = 0$ and $\mathsf{K}_{n+1} = \mathsf{K}_n^+$ if $\mathsf{B}_{n+1} = 1$. Similar relationships hold for $\mathsf{H}_n$ and $\mathsf{Z}_n$.

Let $\hat{\mathsf{K}}_n, \hat{\mathsf{H}}_n$, and $\hat{\mathsf{Z}}_n$ denote the boundary-state-informed versions of $\mathsf{K}_n, \mathsf{Z}_n$, and $\mathsf{H}_n$, respectively. That is,

$$\begin{aligned} \hat{\mathsf{K}}_n &= \mathcal{K}(\mathsf{U}_i|\mathsf{Q}_i, \mathsf{S}_N, \mathsf{S}_0), \\ \hat{\mathsf{Z}}_n &= \mathcal{Z}(\mathsf{U}_i|\mathsf{Q}_i, \mathsf{S}_N, \mathsf{S}_0), \\ \hat{\mathsf{H}}_n &= \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i, \mathsf{S}_N, \mathsf{S}_0), \end{aligned} \tag{25}$$

[7] Let $\mathcal{A}$ and $\mathcal{B}$ be two $\sigma$-algebras. If for any two events $A \in \mathcal{A}$ and $B \in \mathcal{B}$ we have $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)\mathbb{P}(B)$ then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Assume to the contrary that for some events $A_0$, $B_0$, $\mathbb{P}(A_0 \cap B_0) < \mathbb{P}(A_0)\mathbb{P}(B_0)$. Denote the complement of $A_0$ by $\bar{A}_0$. Since $\bar{A}_0 \in \mathcal{A}$, we obtain a contradiction: $\mathbb{P}(B_0) = \mathbb{P}(\bar{A}_0 \cap B_0) + \mathbb{P}(A_0 \cap B_0) < \mathbb{P}(A_0)\mathbb{P}(B_0) + \mathbb{P}(\bar{A}_0)\mathbb{P}(B_0) = \mathbb{P}(B_0)$.

where $i-1=(\mathsf{B}_1\mathsf{B}_2\cdots\mathsf{B}_n)_2$. By (24), $\mathsf{K}_n \le \hat{\mathsf{K}}_n$, $\mathsf{Z}_n \ge \hat{\mathsf{Z}}_n$, and $\mathsf{H}_n \ge \hat{\mathsf{H}}_n$ for any $n$. Similar to (9), we have

$$\hat{\mathsf{K}}_{n+1} = \begin{cases} \mathcal{K}(\mathsf{U}_i + \mathsf{V}_i|\mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_{2N}), & \text{if } \mathsf{B}_{n+1}=0, \\ \mathcal{K}(\mathsf{V}_i|\mathsf{U}_i + \mathsf{V}_i, \mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_{2N}), & \text{if } \mathsf{B}_{n+1}=1. \end{cases} \tag{26}$$

Relationships akin to (26) hold for $\hat{\mathsf{Z}}_{n+1}$ and $\hat{\mathsf{H}}_{n+1}$, with $\mathcal{K}$ replaced with $\mathcal{Z}$ and $\mathcal{H}$, respectively. We use the mnemonic $\hat{\mathsf{K}}_{n+1}^{-} = \mathcal{K}(\mathsf{U}_i + \mathsf{V}_i|\mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_{2N})$ and $\hat{\mathsf{K}}_{n+1}^{+} = \mathcal{K}(\mathsf{V}_i|\mathsf{U}_i + \mathsf{V}_i, \mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_{2N})$, and similar mnemonics for the BSI Bhattachryya and conditional entropy processes.

### *A. Existing Polarization Results for FAIM Processes*

In [12], a class of processes with memory was considered. For this class, the authors showed that the conditional entropy process polarizes and that the Bhattacharyya process polarizes fast to 0.

Specifically, let

$$\mathcal{H}_\star(\mathsf{X}|\mathsf{Y}) \triangleq \lim_{N\to\infty} \frac{1}{N}\mathcal{H}(\mathsf{X}_1^N|\mathsf{Y}_1^N).$$

This limit exists due to stationarity [17, Section 4.2] and the identity $\mathcal{H}(\mathsf{X}_1^N|\mathsf{Y}_1^N) = \mathcal{H}(\mathsf{X}_1^N, \mathsf{Y}_1^N) - \mathcal{H}(\mathsf{Y}_1^N)$.

**Theorem 7.** *[12, Theorems 1,2,5,6] For a strictly stationary $\psi$-mixing process $(\mathsf{X}_j, \mathsf{Y}_j)$, $j \in \mathbb{Z}$, with $\psi(0) < \infty$:*

1) *$\mathsf{H}_n$ polarizes to $\mathsf{H}_\infty$ with $\mathbb{P}(\mathsf{H}_\infty = 1) = \mathcal{H}_\star(\mathsf{X}|\mathsf{Y})$;*
2) *$\mathsf{Z}_n$ polarizes fast to 0 with $\beta < 1/2$.*

*In particular, for any $\epsilon > 0$,*

$$\lim_{N\to\infty} \frac{1}{N}\left|\{i : \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i) > 1-\epsilon\}\right| = \mathcal{H}_\star(\mathsf{X}|\mathsf{Y}), \tag{27a}$$

$$\lim_{N\to\infty} \frac{1}{N}\left|\{i : \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i) < \epsilon\}\right| = 1-\mathcal{H}_\star(\mathsf{X}|\mathsf{Y}), \tag{27b}$$

*and for any $\beta < 1/2$,*

$$\lim_{N\to\infty} \frac{1}{N}\left|\left\{i : \mathcal{Z}(\mathsf{U}_i|\mathsf{Q}_i) < 2^{-N^\beta}\right\}\right| = 1-\mathcal{H}_\star(\mathsf{X}|\mathsf{Y}). \tag{28}$$

To prove Theorem 7, the conditional entropy process $\mathsf{H}_n$ was shown to be a bounded supermartingale, so it converges almost surely to some random variable $\mathsf{H}_\infty$. This latter random variable was shown to be a $\{0,1\}$-random variable with $\mathbb{P}(\mathsf{H}_\infty = 1) = 1 - \mathbb{P}(\mathsf{H}_\infty = 0) = \mathcal{H}_\star(\mathsf{X}|\mathsf{Y})$. This yields (27).

Equation (28) is based on the observation that

$$\mathbb{P}\left(\mathsf{Z}_n < 2^{-N^\beta}\right) = \frac{1}{N}\left|\left\{i : \mathcal{Z}(\mathsf{U}_i|\mathsf{Q}_i) < 2^{-N^\beta}\right\}\right|. \tag{29}$$

First, the Bhattacharyya process $\mathsf{Z}_n$ was also shown to converge almost surely to $\mathsf{H}_\infty$. Next, using the mixing property, the authors showed that $\mathsf{Z}_n^- \le 2\psi(0)\mathsf{Z}_n$ and $\mathsf{Z}_n^+ \le \psi(0)\mathsf{Z}_n^2$. This allowed them to invoke Lemma 3 and obtain (28).

**Corollary 8.** *Let $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$, $j \in \mathbb{Z}$ be a FAIM process. Then,*

1) *Its conditional entropy process $\mathsf{H}_n$ polarizes to $\mathsf{H}_\infty$ with $\mathbb{P}(\mathsf{H}_\infty = 1) = \mathcal{H}_\star(\mathsf{X}|\mathsf{Y})$.*
2) *Its Bhattacharyya process $\mathsf{Z}_n$ polarizes fast to 0 with any $\beta < 1/2$.*

*Proof:* By Lemma 5, blocks of FAIM processes are $\psi$-mixing and satisfy the requirements of Theorem 7. ■

Theorem 7, and consequently Corollary 8, are silent on the rate of polarization of $\mathsf{Z}_n$ to 1. In the sequel we establish a compatible claim for FAIM processes. To do this, we exploit the structure of FAIM processes by calling upon the BSI processes $\hat{\mathsf{H}}_n$ and $\hat{\mathsf{K}}_n$.

### *B. Polarization of the BSI Distribution Parameters*

This section is concerned with proving that the BSI distribution parameters polarize. We achieve this by first showing that the BSI conditional entropy polarizes and then using Lemma 1 to establish polarization of the BSI Bhattacharyya parameter and BSI total variation distance.

**Theorem 9.** *Let $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$, $j \in \mathbb{Z}$ be a FAIM process. The BSI conditional entropy process $\hat{\mathsf{H}}_n$ polarizes to $\hat{\mathsf{H}}_\infty$ and $\hat{\mathsf{H}}_\infty = \mathsf{H}_\infty$ almost surely.*

*In particular, for any $\epsilon > 0$,*

$$\lim_{N\to\infty} \frac{1}{N}\left|\{i : \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i, \mathsf{S}_0, \mathsf{S}_N) > 1-\epsilon\}\right| = \mathcal{H}_\star(\mathsf{X}|\mathsf{Y}),$$

$$\lim_{N\to\infty} \frac{1}{N}\left|\{i : \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i, \mathsf{S}_0, \mathsf{S}_N) < \epsilon\}\right| = 1-\mathcal{H}_\star(\mathsf{X}|\mathsf{Y}).$$

*Proof:* Consider two adjacent blocks of length $N = 2^n$ and let $i-1 = (\mathsf{B}_1\mathsf{B}_2\cdots\mathsf{B}_n)_2$. Recall from (10) that $(\mathsf{U}_i, \mathsf{Q}_i) = f(\mathsf{X}_1^N, \mathsf{Y}_1^N)$ and $(\mathsf{V}_i, \mathsf{R}_i) = f(\mathsf{X}_{N+1}^{2N}, \mathsf{Y}_{N+1}^{2N})$, where the function $f$ depends on the index $i$ (see Figure 2). Using (20b) we obtain

$$P_{\mathsf{U}_i,\mathsf{V}_i|\mathsf{Q}_i,\mathsf{R}_i,\mathsf{S}_0,\mathsf{S}_N,\mathsf{S}_{2N}} = P_{\mathsf{U}_i|\mathsf{Q}_i,\mathsf{S}_0,\mathsf{S}_N} \cdot P_{\mathsf{V}_i|\mathsf{R}_i,\mathsf{S}_N,\mathsf{S}_{2N}}. \tag{30}$$

Thus,

$$\begin{aligned}
\hat{\mathsf{H}}_n &\overset{(a)}{=} \frac{1}{2}\Big(\mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i, \mathsf{S}_0, \mathsf{S}_N) + \mathcal{H}(\mathsf{V}_i|\mathsf{R}_i, \mathsf{S}_N, \mathsf{S}_{2N})\Big) \\
&\overset{(b)}{=} \frac{1}{2}\mathcal{H}(\mathsf{U}_i, \mathsf{V}_i|\mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_N, \mathsf{S}_{2N}) \\
&\overset{(c)}{=} \frac{1}{2}\mathcal{H}(\mathsf{U}_i + \mathsf{V}_i, \mathsf{V}_i|\mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_N, \mathsf{S}_{2N}) \\
&\overset{(d)}{=} \frac{1}{2}\Big(\mathcal{H}(\mathsf{U}_i + \mathsf{V}_i|\mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_N, \mathsf{S}_{2N}) \\
&\qquad + \mathcal{H}(\mathsf{V}_i|\mathsf{U}_i + \mathsf{V}_i, \mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_N, \mathsf{S}_{2N})\Big) \\
&\overset{(e)}{\le} \frac{1}{2}\Big(\mathcal{H}(\mathsf{U}_i + \mathsf{V}_i|\mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_{2N}) \\
&\qquad + \mathcal{H}(\mathsf{V}_i|\mathsf{U}_i + \mathsf{V}_i, \mathsf{Q}_i, \mathsf{R}_i, \mathsf{S}_0, \mathsf{S}_{2N})\Big) \\
&= \frac{1}{2}\Big(\hat{\mathsf{H}}_n^- + \hat{\mathsf{H}}_n^+\Big),
\end{aligned}$$

where (a) is by stationarity, (b) is by (30), (c) is because the mapping $(\mathsf{U}, \mathsf{V}) \mapsto (\mathsf{U}+\mathsf{V}, \mathsf{V})$ is one-to-one and onto, (d) is by the chain rule for entropies, and (e) is by (6c).

By (26) (applied to the BSI conditional entropy), $\hat{\mathsf{H}}_n$ is a submartingale sequence:

$$\frac{1}{2}\Big(\hat{\mathsf{H}}_n^- + \hat{\mathsf{H}}_n^+\Big) = \mathbb{E}\left[\hat{\mathsf{H}}_{n+1}\middle|\hat{\mathsf{H}}_n, \hat{\mathsf{H}}_{n-1}, \ldots, \hat{\mathsf{H}}_1\right] \ge \hat{\mathsf{H}}_n.$$

It is also bounded, as $\hat{\mathsf{H}}_n \in [0,1]$ for any $n$. Thus, it converges almost surely to some random variable $\hat{\mathsf{H}}_\infty \in [0,1]$, [22, Theorem 35.4].

Denote $\Delta\mathsf{H}_n = \mathsf{H}_n - \hat{\mathsf{H}}_n$. The sequence $\Delta\mathsf{H}_n$ converges almost surely to the random variable $\Delta\mathsf{H}_\infty = \mathsf{H}_\infty - \hat{\mathsf{H}}_\infty$. This is because $\hat{\mathsf{H}}_n$ converges almost surely to $\hat{\mathsf{H}}_\infty$, and, by Corollary 8, $\mathsf{H}_n$ converges almost surely to $\mathsf{H}_\infty$. By (6), $\Delta\mathsf{H}_n \geq 0$ for any $n$, which implies that $\Delta\mathsf{H}_\infty \geq 0$ almost surely. We now show that $\Delta\mathsf{H}_\infty = 0$ almost surely. To this end, we will need the following lemma, whose proof is postponed to the end of this theorem.

**Lemma 10.** *The sequence* $\Delta\mathsf{H}_n$ *satisfies*

$$\lim_{n\to\infty} \mathbb{E}\left[\Delta\mathsf{H}_n\right] = 0.$$

Since $\Delta\mathsf{H}_n$ converges to $\Delta\mathsf{H}_\infty$ almost surely, we specifically have $\liminf_{n\to\infty} \Delta\mathsf{H}_n = \Delta\mathsf{H}_\infty$ almost surely. Using Fatou's lemma[8] for the non-negative sequence $\Delta\mathsf{H}_n$, $n = 1, 2, \ldots$ we obtain

$$\begin{aligned} 0 \leq \mathbb{E}\left[\Delta\mathsf{H}_\infty\right] &= \mathbb{E}\left[\liminf_{n\to\infty} \Delta\mathsf{H}_n\right] \\ &\leq \liminf_{n\to\infty} \mathbb{E}\left[\Delta\mathsf{H}_n\right] = \lim_{n\to\infty} \mathbb{E}\left[\Delta\mathsf{H}_n\right] = 0. \end{aligned}$$

Thus, $\mathbb{E}\left[\Delta\mathsf{H}_\infty\right] = 0$. By Markov's inequality, $\mathbb{P}(\Delta\mathsf{H}_\infty \geq \delta) \leq \mathbb{E}\left[\Delta\mathsf{H}_\infty\right]/\delta = 0$ for any $\delta > 0$; consequently, $\mathbb{P}(\Delta\mathsf{H}_\infty = 0) = \mathbb{P}(\mathsf{H}_\infty = \hat{\mathsf{H}}_\infty) = 1$. Put another way, $\hat{\mathsf{H}}_\infty = \mathsf{H}_\infty$ almost surely.

Recall that $\mathsf{H}_\infty$ is a $\{0, 1\}$ random variable with $\mathbb{P}(\mathsf{H}_\infty = 1) = \mathcal{H}_\star(\mathsf{X}|\mathsf{Y})$. Since $\hat{\mathsf{H}}_\infty = \mathsf{H}_\infty$ almost surely, and

$$\begin{aligned} \mathbb{P}\left(\hat{\mathsf{H}}_n > 1-\epsilon\right) &= \frac{1}{N}\left|\{i : \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i, \mathsf{S}_0, \mathsf{S}_N) > 1-\epsilon\}\right|, \\ \mathbb{P}\left(\hat{\mathsf{H}}_n < \epsilon\right) &= \frac{1}{N}\left|\{i : \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i, \mathsf{S}_0, \mathsf{S}_N) < \epsilon\}\right|, \end{aligned}$$

the proof is complete. ■

*Proof of Lemma 10:* By (6), $\Delta\mathsf{H}_n \geq 0$, so $\mathbb{E}\left[\Delta\mathsf{H}_n\right] \geq 0$ as well.

Using the chain rule for conditional entropies and since the transformation $\mathsf{U}_1^N = \mathsf{X}_1^N G_N$ is one-to-one and onto,

$$\mathbb{E}\left[\mathsf{H}_n\right] = \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(\mathsf{U}_i|\mathsf{Q}_i) = \frac{\mathcal{H}(\mathsf{U}_1^N|\mathsf{Y}_1^N)}{N} = \frac{\mathcal{H}(\mathsf{X}_1^N|\mathsf{Y}_1^N)}{N}.$$

Similarly, $\mathbb{E}\left[\hat{\mathsf{H}}_n\right] = \mathcal{H}(\mathsf{X}_1^N|\mathsf{Y}_1^N, \mathsf{S}_0, \mathsf{S}_N)/N$. Thus,

$$\begin{aligned} \mathbb{E}\left[\Delta\mathsf{H}_n\right] &= \frac{1}{N}\left(\mathcal{H}(\mathsf{X}_1^N|\mathsf{Y}_1^N) - \mathcal{H}(\mathsf{X}_1^N|\mathsf{Y}_1^N, \mathsf{S}_0, \mathsf{S}_N)\right) \\ &\overset{(a)}{=} \frac{1}{N}\left(\mathcal{H}(\mathsf{S}_0, \mathsf{S}_N|\mathsf{Y}_1^N) - \mathcal{H}(\mathsf{S}_0, \mathsf{S}_N|\mathsf{X}_1^N, \mathsf{Y}_1^N)\right) \\ &\overset{(b)}{\leq} \frac{2\log_2(|\mathcal{S}|)}{N}. \end{aligned}$$

To see (a), note that for any 3 random variables $\mathsf{A}, \mathsf{B}, \mathsf{C}$ we have $\mathcal{H}(\mathsf{A}, \mathsf{B}|\mathsf{C}) = \mathcal{H}(\mathsf{A}|\mathsf{C}) + \mathcal{H}(\mathsf{B}|\mathsf{A}, \mathsf{C}) = \mathcal{H}(\mathsf{B}|\mathsf{C}) + \mathcal{H}(\mathsf{A}|\mathsf{B}, \mathsf{C})$. Rearranging and setting $\mathsf{A} = \mathsf{X}_1^N$, $\mathsf{B} = (\mathsf{S}_0, \mathsf{S}_N)$ and $\mathsf{C} = \mathsf{Y}_1^N$ yields (a). Inequality (b) is since $\mathsf{S}_0, \mathsf{S}_N$ take values in the finite alphabet $\mathcal{S}$ and the conditional entropy is non-negative.

Combining these inequalities, and recalling that $N = 2^n$, we obtain

$$0 \leq \mathbb{E}\left[\Delta\mathsf{H}_n\right] \leq 2\log_2(|\mathcal{S}|)/2^n.$$

This holds for any $n$. We take limits and use the sandwich rule to yield $\lim_{n\to\infty} \mathbb{E}\left[\Delta\mathsf{H}_n\right] = 0$, as desired. ■

The following corollary is a direct consequence of the definition of almost-sure convergence, Lemma 1, Corollary 8, and Theorem 9.

**Corollary 11.**

1) *The sequences* $\mathsf{Z}_n$ *and* $\hat{\mathsf{Z}}_n$ *polarize to random variables* $\mathsf{Z}_\infty$ *and* $\hat{\mathsf{Z}}_\infty$*, respectively. Moreover,* $\mathsf{Z}_\infty = \hat{\mathsf{Z}}_\infty = \mathsf{H}_\infty$ *almost surely.*
2) *The sequences* $\mathsf{K}_n$ *and* $\hat{\mathsf{K}}_n$ *polarize to random variables* $\mathsf{K}_\infty$ *and* $\hat{\mathsf{K}}_\infty$*, respectively. Moreover,* $\mathsf{K}_\infty = \hat{\mathsf{K}}_\infty = 1 - \mathsf{H}_\infty$ *almost surely.*

*Proof:* The proofs of both items are essentially the same, so we prove only the first item.

Recall the definition of almost-sure convergence of a sequence of random variables. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathsf{A}, \mathsf{A}_1, \mathsf{A}_2, \ldots$ be a sequence of $\mathcal{F}$-measurable random variables defined on this space. A random variable is a deterministic function from $\Omega$ to $\mathbb{R}$. We say that $\mathsf{A}_n$ converges to $\mathsf{A}$ almost surely if the set

$$A = \left\{\omega \in \Omega : \lim_{n\to\infty} \mathsf{A}_n(\omega) = \mathsf{A}(\omega)\right\}$$

satisfies $\mathbb{P}(A) = 1$.

Now, let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space in which $\mathsf{H}_n, \hat{\mathsf{H}}_n, \mathsf{Z}_n, \hat{\mathsf{Z}}_n$, $n = 1, 2, \ldots$ as well as $\mathsf{H}_\infty$ and $\hat{\mathsf{H}}_\infty$ are defined.

By Corollary 8 and Theorem 9, $\mathsf{H}_n$ and $\hat{\mathsf{H}}_n$ converge almost surely to $\mathsf{H}_\infty$ and $\hat{\mathsf{H}}_\infty$, respectively, and $\mathsf{H}_\infty = \hat{\mathsf{H}}_\infty$ almost surely. Thus, we denote

$$H = \left\{\omega \in \Omega : \lim_{n\to\infty} \mathsf{H}_n(\omega) = \lim_{n\to\infty} \hat{\mathsf{H}}_n(\omega) = \mathsf{H}_\infty(\omega)\right\}.$$

By definition of almost sure convergence, $\mathbb{P}(H) = 1$.

Since $\mathsf{H}_\infty(\omega) \in \{0, 1\}$ almost surely, we split $H = H_0 \uplus H_1 \uplus H_\emptyset$, such that $\mathsf{H}_\infty(\omega) = 0$ for any $\omega \in H_0$; $\mathsf{H}_\infty(\omega) = 1$ for any $\omega \in H_1$; and $H_\emptyset$ is a set of measure zero. By Lemma 1, we have $\mathsf{H}_n(\omega) \leq \mathsf{Z}_n(\omega) \leq \sqrt{\mathsf{H}_n(\omega)}$ for any $\omega$. Thus, $\lim_{n\to\infty} \mathsf{Z}_n(\omega) = 0$ for all $\omega \in H_0$ and $\lim_{n\to\infty} \mathsf{Z}_n(\omega) = 1$ for all $\omega \in H_1$. We conclude that $\mathsf{Z}_n$ converges almost surely to a $\{0, 1\}$-random variable $\mathsf{Z}_\infty$ and $\mathsf{Z}_\infty = \mathsf{H}_\infty$ almost surely. Using similar arguments, $\hat{\mathsf{Z}}_n$ converges almost surely to a random variable $\hat{\mathsf{Z}}_\infty$ and $\hat{\mathsf{Z}}_\infty = \hat{\mathsf{H}}_\infty$ almost surely. By Theorem 9, $\hat{\mathsf{H}}_\infty = \mathsf{H}_\infty$ almost surely. ■

### *C. Fast Polarization of the Bhattacharyya Process to* 1

In this section, we prove that the Bhattacharyya process $\mathsf{Z}_n$ of a FAIM process polarizes fast to 1.

Theorem 13, the main theorem of this section, relies on an inequality akin to (13) for the BSI total variation distance. We state the inequality in Proposition 12, and postpone its proof to the end of the section.

**Proposition 12.** *Let* $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$*,* $j \in \mathbb{Z}$ *be a FAIM process. Then,*

$$\hat{\mathsf{K}}_{n+1} \leq \begin{cases} \psi(0)\hat{\mathsf{K}}_n^2, & \text{if } \mathsf{B}_{n+1} = 0, \\ 2\hat{\mathsf{K}}_n, & \text{if } \mathsf{B}_{n+1} = 1. \end{cases} \tag{31}$$

[8] Fatou's lemma [22, Theorem 16.3] states that if $\mathsf{A}_n$, $n = 1, 2, \ldots$ is a sequence of non-negative random variables then $\mathbb{E}\left[\liminf_{n\to\infty} \mathsf{A}_n\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[\mathsf{A}_n\right]$.

Here, $\psi(0)$ is as defined in (19), i.e.,

$$\psi(0) = \max_{a} \frac{1}{\pi_0(a)} = \max_{b} \frac{1}{\pi_N(b)} \geq 1. \tag{32}$$

Since the state sequence is stationary, finite-state, aperiodic, and irreducible, $\psi(0) < \infty$.

**Theorem 13.** *Let $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$, $j \in \mathbb{Z}$ be a FAIM process. Then $\mathsf{Z}_n$ polarizes fast to* 1 *and for any $\beta < 1/2$,*

$$\lim_{N\to\infty} \frac{1}{N} \left| \left\{ i : \mathcal{Z}(\mathsf{U}_i|\mathsf{Q}_i) > 1 - 2^{-N^{\beta}} \right\} \right| = \mathcal{H}_{\star}(\mathsf{X}|\mathsf{Y}). \tag{33}$$

*Proof:* Fix $\beta < 1/2$. By Corollary 11 and (31), we can invoke Lemma 3 for $\hat{\mathsf{K}}_n$ with $E = 1/2$. Consequently, $\hat{\mathsf{K}}_n$ polarizes fast to 0, i.e.,

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}\left(\hat{\mathsf{K}}_n < 2^{-N^{\beta}}\right) &= \mathbb{P}\left(\hat{\mathsf{K}}_{\infty} = 0\right) \\ &= \mathbb{P}\left(\mathsf{H}_{\infty} = 1\right) = \mathcal{H}_{\star}(\mathsf{X}|\mathsf{Y}). \end{aligned}$$

For any $n$, by (4a), (4b), and (24),

$$1 - \mathsf{Z}_n \leq 1 - \mathsf{H}_n \leq \mathsf{K}_n \leq \hat{\mathsf{K}}_n.$$

Thus,

$$\mathbb{P}\left(\mathsf{Z}_n > 1 - 2^{-N^{\beta}}\right) \geq \mathbb{P}\left(\hat{\mathsf{K}}_n < 2^{-N^{\beta}}\right).$$

Taking limits, we obtain that

$$\liminf_{n\to\infty} \mathbb{P}(\mathsf{Z}_n > 1 - 2^{-N^{\beta}}) \geq \mathcal{H}_{\star}(\mathsf{X}|\mathsf{Y}).$$

On the other hand, by Corollary 8,

$$\lim_{n\to\infty} \mathbb{P}(\mathsf{Z}_n < 2^{-N^{\beta}}) = 1 - \mathcal{H}_{\star}(\mathsf{X}|\mathsf{Y}).$$

Recalling that $\mathbb{P}(\mathsf{Z}_n < 2^{-N^{\beta}}) + \mathbb{P}(\mathsf{Z}_n > 1 - 2^{-N^{\beta}}) \leq 1$ for any $n$, we take limits to obtain $\limsup_{n\to\infty} \mathbb{P}(\mathsf{Z}_n > 1 - 2^{-N^{\beta}}) \leq \mathcal{H}_{\star}(\mathsf{X}|\mathsf{Y})$. Therefore, we conclude that

$$\lim_{n\to\infty} \mathbb{P}\left(\mathsf{Z}_n > 1 - 2^{-N^{\beta}}\right) = \mathcal{H}_{\star}(\mathsf{X}|\mathsf{Y}).$$

To obtain (33), note that by definition of the Bhattacharyya process,

$$\mathbb{P}\left(\mathsf{Z}_n > 1 - 2^{-N^{\beta}}\right) = \frac{1}{N} \left| \left\{ i : \mathcal{Z}(\mathsf{U}_i|\mathsf{Q}_i) > 1 - 2^{-N^{\beta}} \right\} \right|.$$

Taking limits completes the proof. ■

*Proof of Proposition 12:* The proof follows along the lines of the proof of Proposition 4.

Consider two adjacent blocks of length $N = 2^n$ and let $i - 1 = (\mathsf{B}_1\mathsf{B}_2 \cdots \mathsf{B}_n)_2$. This is illustrated in Figure 2. Recall from (10) that there is a function $f$ that depends on $i$ such that $(\mathsf{U}_i, \mathsf{Q}_i) = f(\mathsf{X}_1^N, \mathsf{Y}_1^N)$ and $(\mathsf{V}_i, \mathsf{R}_i) = f(\mathsf{X}_{N+1}^{2N}, \mathsf{Y}_{N+1}^{2N})$. By stationarity,

$$\hat{\mathsf{K}}_n = \sum_{a,b\in\mathcal{S}} \pi_{N,0}(b,a)\mathcal{K}_a^b(\mathsf{U}_i|\mathsf{Q}_i) = \sum_{b,c\in\mathcal{S}} \pi_{2N,N}(c,b)\mathcal{K}_b^c(\mathsf{V}_i|\mathsf{R}_i). \tag{34}$$

As in (21), we denote

$$P_a^c(u,q) = P_{\mathsf{U}_i,\mathsf{Q}_i|\mathsf{S}_N,\mathsf{S}_0}(u,q|c,a) = P_{\mathsf{V}_i,\mathsf{R}_i|\mathsf{S}_{2N},\mathsf{S}_N}(u,q|c,a).$$

The right-most equality is due to stationarity. We further denote $P_a^c(s) = P_a^c(0,s) + P_a^c(1,s)$; in particular, $\sum_s P_a^c(s) = 1$.

Denote

$$\begin{aligned} \mu(b) &= \pi_{2N|N}(c|b)\pi_{N|0}(b|a)\pi_0(a) \\ &= \frac{\pi_{2N,N}(c,b) \cdot \pi_{N,0}(b,a)}{\pi_N(b)}. \end{aligned}$$

We deliberately omitted the dependence on $a, c$ from this notation to simplify the expressions that follow. Observe that by (32),

$$\mu(b) \leq \psi(0) \cdot \pi_{2N,N}(c,b) \cdot \pi_{N,0}(b,a). \tag{35}$$

Also, since $\pi_N(b) = \sum_{a\in\mathcal{S}} \pi_{N,0}(b,a) = \sum_{c\in\mathcal{S}} \pi_{2N,N}(c,b)$, we have

$$\sum_{a\in\mathcal{S}} \mu(b) = \pi_{2N,N}(c,b), \quad \sum_{c\in\mathcal{S}} \mu(b) = \pi_{N,0}(b,a). \tag{36}$$

By (16) and (21),

$$\begin{aligned} &\pi_{2N,0}(c,a)P_{\mathsf{U}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i|\mathsf{S}_{2N},\mathsf{S}_0}(u,v,q,r|c,a) \\ &\quad = \pi_0(a)\pi_{2N|0}(c|a)P_{\mathsf{U}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i|\mathsf{S}_{2N},\mathsf{S}_0}(u,v,q,r|c,a) \\ &\quad = \pi_0(a)P_{\mathsf{U}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i,\mathsf{S}_{2N}|\mathsf{S}_0}(u,v,q,r,c|a) \\ &\quad = \pi_0(a)\sum_{b\in\mathcal{S}} P_{\mathsf{U}_i,\mathsf{Q}_i,\mathsf{S}_N|\mathsf{S}_0}(u,q,b|a)P_{\mathsf{V}_i,\mathsf{R}_i,\mathsf{S}_{2N}|\mathsf{S}_N}(v,r,c|b) \\ &\quad = \pi_0(a)\sum_{b\in\mathcal{S}} \pi_{N|0}(b|a)P_a^b(u,q)\pi_{2N|N}(c|b)P_b^c(v,r) \\ &\quad = \sum_{b\in\mathcal{S}} \mu(b)P_a^b(u,q)P_b^c(v,r). \end{aligned} \tag{37}$$

Set $\mathsf{T}_i = \mathsf{U}_i + \mathsf{V}_i$. Using (9), a single-step polarization from $\hat{\mathsf{K}}_n$ to $\hat{\mathsf{K}}_{n+1}$ becomes

$$\hat{\mathsf{K}}_{n+1} = \begin{cases} \displaystyle\sum_{a,c\in\mathcal{S}} \pi_{2N,0}(c,a)\mathcal{K}_a^c(\mathsf{T}_i|\mathsf{Q}_i,\mathsf{R}_i), & \text{if } \mathsf{B}_{n+1} = 0, \\ \displaystyle\sum_{a,c\in\mathcal{S}} \pi_{2N,0}(c,a)\mathcal{K}_a^c(\mathsf{V}_i|\mathsf{T}_i,\mathsf{Q}_i,\mathsf{R}_i), & \text{if } \mathsf{B}_{n+1} = 1. \end{cases}$$

Here, $\mathcal{K}_a^c(\mathsf{T}_i|\mathsf{Q}_i,\mathsf{R}_i)$ and $\mathcal{K}_a^c(\mathsf{V}_i|\mathsf{T}_i,\mathsf{Q}_i,\mathsf{R}_i)$ are computed as in (23), only for a block of length $2N$ with initial state $\mathsf{S}_0 = a$ and final state $\mathsf{S}_{2N} = c$. At the middle of the block we have state $\mathsf{S}_N = b$. Using (11), we denote

$$\begin{aligned} \bar{P}_a^c(t,v,q,r) &= P_{\mathsf{T}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i|\mathsf{S}_{2N},\mathsf{S}_0}(t,v,q,r|c,a) \\ &= P_{\mathsf{U}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i|\mathsf{S}_{2N},\mathsf{S}_0}(t+v,v,q,r|c,a) \end{aligned} \tag{38}$$

and

$$\begin{aligned} \bar{P}_a^c(t,q,r) &= P_{\mathsf{T}_i,\mathsf{Q}_i,\mathsf{R}_i|\mathsf{S}_{2N},\mathsf{S}_0}(t,q,r|c,a) \\ &= \textstyle\sum_{v=0}^{1} \bar{P}_a^c(t,v,q,r). \end{aligned} \tag{39}$$

Consider first the case $\mathsf{B}_{n+1} = 0$:

$$\begin{aligned}
&\pi_{2N,0}(c,a)\mathcal{K}_a^c(\mathsf{T}_i|\mathsf{Q}_i,\mathsf{R}_i)\\
&= \pi_{2N,0}(c,a)\sum_{q,r}\left|\bar{P}_a^c(0,q,r) - \bar{P}_a^c(1,q,r)\right|\\
&= \sum_{q,r}\left|\pi_{2N,0}(c,a)\bar{P}_a^c(0,q,r) - \pi_{2N,0}(c,a)\bar{P}_a^c(1,q,r)\right|\\
&\overset{(a)}{=} \sum_{q,r}\left|\sum_{b\in\mathcal{S}}\mu(b)\sum_{v=0}^{1}P_b^c(v,r)(P_a^b(v,q) - P_a^b(v+1,q))\right|\\
&\overset{(b)}{\le} \sum_{\substack{q,r,\\ b\in\mathcal{S}}}\mu(b)\left|\sum_{v=0}^{1}P_b^c(v,r)(P_a^b(v,q) - P_a^b(v+1,q))\right|\\
&= \sum_{\substack{q,r,\\ b\in\mathcal{S}}}\mu(b)\left|P_a^b(0,q) - P_a^b(1,q)\right|\cdot\left|P_b^c(0,r) - P_b^c(1,r)\right|\\
&= \sum_{b\in\mathcal{S}}\mu(b)\mathcal{K}_a^b(\mathsf{U}_i|\mathsf{Q}_i)\mathcal{K}_b^c(\mathsf{V}_i|\mathsf{R}_i)\\
&\overset{(c)}{\le} \psi(0)\sum_{b\in\mathcal{S}}\left(\pi_{2N,N}(c,b)\mathcal{K}_b^c(\mathsf{V}_i|\mathsf{R}_i)\right)\cdot\left(\pi_{N,0}(b,a)\mathcal{K}_a^b(\mathsf{U}_i|\mathsf{Q}_i)\right)\\
&\overset{(d)}{\le} \psi(0)\sum_{b\in\mathcal{S}}\pi_{2N,N}(c,b)\mathcal{K}_b^c(\mathsf{V}_i|\mathsf{R}_i)\sum_{b'\in\mathcal{S}}\pi_{N,0}(b',a)\mathcal{K}_a^{b'}(\mathsf{U}_i|\mathsf{Q}_i),
\end{aligned}$$

where (a) first expands $\bar{P}_a^c(0,q,r)$ and $\bar{P}_a^c(1,q,r)$ according to (39) and then (38), and finally applies (37); (b) is by the triangle inequality; (c) is by (35); and (d) is by the inequality $\sum_j a_j b_j \le \sum_j a_j \sum_{j'} b_{j'}$, which holds for $a_j, b_j \ge 0$. By (34), the sum over $a, c \in \mathcal{S}$ yields

$$\sum_{a,c\in\mathcal{S}}\pi_{2N,0}(c,a)\mathcal{K}_a^c(\mathsf{T}_i|\mathsf{Q}_i,\mathsf{R}_i) \le \psi(0)\hat{\mathsf{K}}_n^2.$$

Next, let $\mathsf{B}_{n+1} = 1$. We have

$$\begin{aligned}
&\pi_{2N,0}(c,a)\mathcal{K}_a^c(\mathsf{V}_i|\mathsf{T}_i,\mathsf{Q}_i,\mathsf{R}_i)\\
&\quad= \pi_{2N,0}(c,a)\sum_{t,q,r}\left|\bar{P}_a^c(t,0,q,r) - \bar{P}_a^c(t,1,q,r)\right|\\
&\quad= \sum_{t,q,r}\left|\pi_{2N,0}(c,a)\bar{P}_a^c(t,0,q,r) - \pi_{2N,0}(c,a)\bar{P}_a^c(t,1,q,r)\right|\\
&\quad\overset{(a)}{=} \sum_{t,q,r}\left|\sum_{b\in\mathcal{S}}\mu(b)(P_a^b(t,q)P_b^c(0,r) - P_s^b(t+1,q)P_b^c(1,r))\right|\\
&\quad\overset{(b)}{=} \frac{1}{2}\sum_{t,q,r}\Bigg|\sum_{b\in\mathcal{S}}\mu(b)P_a^b(q)(P_b^c(0,r) - P_b^c(1,r))\\
&\qquad\qquad + \sum_{b\in\mathcal{S}}\mu(b)P_b^c(r)(P_a^b(t,q) - P_a^b(t+1,q))\Bigg|\\
&\quad\overset{(c)}{\le} \sum_{\substack{q,\\ b\in\mathcal{S}}}\mu(b)P_a^b(q)\left(\sum_r\left|P_b^c(0,r) - P_b^c(1,r)\right|\right)\\
&\qquad + \sum_{\substack{r,\\ b\in\mathcal{S}}}\mu(b)P_b^c(r)\left(\sum_q\left|P_a^b(0,q) - P_a^b(1,q)\right|\right)\\
&\quad= \sum_{b\in\mathcal{S}}\mu(b)\mathcal{K}_b^c(\mathsf{V}_i|\mathsf{R}_i) + \sum_{b\in\mathcal{S}}\mu(b)\mathcal{K}_a^b(\mathsf{U}_i|\mathsf{Q}_i),
\end{aligned}$$

where (a) first expands $\bar{P}_a^c(t,0,q,r)$ and $\bar{P}_a^c(t,1,q,r)$ according to (38), and then applies (37); (b) is by (15); and (c) is by the triangle inequality. Since $\mu(b)$ depends on $a, c$, we use (36) to obtain

$$\begin{aligned}
\sum_{a,b,c\in\mathcal{S}}\mu(b)\mathcal{K}_b^c(\mathsf{V}_i|\mathsf{R}_i) &= \sum_{b,c\in\mathcal{S}}\pi_{2N,N}(c,b)\mathcal{K}_b^c(\mathsf{V}_i|\mathsf{R}_i) = \hat{\mathsf{K}}_n,\\
\sum_{a,b,c\in\mathcal{S}}\mu(b)\mathcal{K}_a^b(\mathsf{U}_i|\mathsf{Q}_i) &= \sum_{a,b\in\mathcal{S}}\pi_{N,0}(b,a)\mathcal{K}_a^b(\mathsf{U}_i|\mathsf{Q}_i) = \hat{\mathsf{K}}_n.
\end{aligned}$$

Thus,

$$\sum_{a,c\in\mathcal{S}}\pi_{2N,0}(c,a)\mathcal{K}_a^c(\mathsf{V}_i|\mathsf{T}_i,\mathsf{Q}_i,\mathsf{R}_i) \le 2\hat{\mathsf{K}}_n.$$

This completes the proof. ■

### *D. Fast Polarization of the BSI Bhattacharyya Process*

Fast polarization of the Bhattacharyya process was established in Corollary 8 and Theorem 13. Implicitly, however, we have also obtained fast polarization of the BSI-Bhattacharyya process $\hat{\mathsf{Z}}_n$, both to 0 and 1. We now make this explicit.

**Corollary 14.** *Let $(\mathsf{X}_j, \mathsf{Y}_j, \mathsf{S}_j)$, $j \in \mathbb{Z}$ be a FAIM process. Then $\hat{\mathsf{Z}}_n$ polarizes fast both to 0 and to 1 with any $\beta < 1/2$.*

*Proof:* Polarization of $\hat{\mathsf{Z}}_n$ was obtained directly in Corollary 11. By (24), $\mathsf{Z}_n \ge \hat{\mathsf{Z}}_n$. Since $\mathsf{Z}_n$ polarizes fast to 0 with any $\beta < 1/2$, so must $\hat{\mathsf{Z}}_n$. We obtain fast polarization of $\hat{\mathsf{Z}}_n$ to 1 by replacing the Bhattacharyya parameter with its BSI counterpart in the proof of Theorem 13. ■

## Acknowledgment

The authors are grateful to Dr. Junya Honda for helpful discussions. We thank the anonymous reviewers for their valuable comments and suggestions, and especially for catching an error in a previous version of Lemma 1.

## Appendix A
## Auxiliary Proofs for Section III

For $\theta \in [0, 1/2]$ we denote

$$\begin{aligned}
k(\theta) &= |\theta - (1-\theta)| = 1 - 2\theta,\\
h(\theta) &= -\theta\log_2\theta - (1-\theta)\log_2(1-\theta),\\
z(\theta) &= 2\sqrt{\theta(1-\theta)}.
\end{aligned}$$

We will need the following lemmas.

**Lemma 15.** *For $\theta \in [0, 1/2]$, we have $z^2(\theta) \le h(\theta) \le z(\theta)$.*

*Proof:* We plot $z^2(\theta)$, $h(\theta)$, and $z(\theta)$ in Figure 3; indeed $z^2(\theta) \le h(\theta) \le z(\theta)$ for $0 \le \theta \le 1/2$. We now prove this formally.

The left-most inequality is obvious for $\theta = 0$. Next, observe that $h(\theta)/\theta$ is convex-∪ in $(0, 1/2]$. To see this, we turn to its second order derivative:

$$\left(\frac{h(\theta)}{\theta}\right)'' = \frac{-(\theta + 2(1-\theta)\ln(1-\theta))}{(1-\theta)\theta^3\ln 2}.$$

We claim that it is nonnegative for $\theta \in (0, 1/2]$, which will imply that $h(\theta)/\theta$ is indeed convex-∪ in $(0, 1/2]$. The denominator is nonnegative, so it remains to show that the numerator is nonnegative as well. Negating the numerator yields $\tau(\theta) = \theta + 2(1-\theta)\ln(1-\theta)$, which is convex-∪ in $[0, 1/2]$ as it is a positive sum of two convex-∪ functions. Since $\tau(0) = 0$ and $\tau(1/2) = 1/2 - \ln 2 < 0$, by definition of convexity,

$$\begin{aligned}\tau(\theta) &= \tau((1-2\theta)\cdot 0 + 2\theta\cdot 1/2)\\ &\le (1-2\theta)\cdot\tau(0) + 2\theta\cdot\tau(1/2)\\ &< 0\end{aligned}$$

for any $\theta \in (0, 1/2]$. This implies that the numerator of the second-order derivative is nonnegative, establishing convexity of $h(\theta)/\theta$.

Consequently, $h(\theta)/\theta$ satisfies the gradient inequality ([25, Theorem 7.6]) by which

$$\begin{aligned}\frac{h(\theta)}{\theta} &\ge \frac{h(1/2)}{1/2} + \left(\frac{h(\theta)}{\theta}\right)'\Bigg|_{\theta=1/2}(\theta - 1/2)\\ &= 2 - 4(\theta - 1/2)\\ &= 4(1-\theta).\end{aligned}$$

This holds for any $\theta \in (0, 1/2]$. Rearranging yields $h(\theta) \ge 4\theta(1-\theta) = z^2(\theta)$, which holds for any $\theta \in [0, 1/2]$.

For the right-most inequality, denote $g(\theta) = h(\theta) - z(\theta)$. Since $g(0) = g(1/2) = 0$, it suffices to show that $g(\theta)$ has a single stationary point in $(0, 1/2)$, and that this point is a minimum.

The stationary points of $g(\theta)$ are the zeros of its derivative

$$g'(\theta) = \log_2\left(\frac{1-\theta}{\theta}\right) - \frac{1-2\theta}{\sqrt{\theta(1-\theta)}}.$$

Recalling that $\theta \in [0, 1/2]$,

$$g'''(\theta) = \frac{(1-2\theta)(4\sqrt{\theta(1-\theta)} - \ln 8)}{4\left(\sqrt{\theta(1-\theta)}\right)^5 \ln 2} \le 0,$$

since

$$4\sqrt{\theta(1-\theta)} - \ln 8 < 4\sqrt{\theta(1-\theta)} - 2 \le 0.$$

Hence, $g'(\theta)$ is concave-∩ in $[0, 1/2]$. Observe that $g'(1/2) = 0$ and $\lim_{\theta\to 0} g'(\theta) = -\infty$, so $g'(\theta)$ can assume the value 0 for at most one point in $(0, 1/2)$. Assume to the contrary that $g'(\theta) < 0$ for all $\theta \in (0, 1/2)$. Then, $g(\theta)$ has no stationary points in $(0, 1/2)$, which, by the mean value theorem, contradicts $g(0) = g(1/2) = 0$. We conclude that $g'(\theta_0) = 0$ for some $\theta_0 \in (0, 1/2)$. Consequently, $\theta_0$ is a stationary point of $g(\theta)$. Since $g'(\theta)$ is concave-∩ and $g'(1/2) = g'(\theta_0) = 0$, then $g'(\theta) > g'(\theta_0)$ for $\theta_0 < \theta < 1/2$ and $g'(\theta) < g'(\theta_0)$ for $0 < \theta < \theta_0$. This implies that $g(\theta) \ge g(\theta_0)$ for any $\theta \in [0, 1/2]$; i.e., $\theta_0$ is the single minimum of $g(\theta)$ in $[0, 1/2]$. ∎

**Lemma 16.** *For $\theta \in [0, 1/2]$, we have $k(\theta) + h(\theta) \ge 1$.*

*Proof:* Both $k(\theta)$ and $h(\theta)$ are continuous and concave-∩ functions in $[0, 1/2]$. Therefore, $\eta(\theta) = k(\theta) + h(\theta)$ is also concave-∩ in this region. Observe that $\eta(0) = \eta(1/2) = 1$. Any $\theta \in [0, 1/2]$ can be written as a convex combination of 0 and 1/2, since $\theta = (1-2\theta)\cdot 0 + (2\theta)\cdot(1/2)$. Thus, by definition of concavity, $\eta(\theta) \ge (1-2\theta)\eta(0) + (2\theta)\eta(1/2) = 1$ for any $\theta \in [0, 1/2]$. ∎

*Proof of Lemma 1:* For any $q$, denote

$$\theta = \theta(q) = \min\{P_{\mathsf{U}|\mathsf{Q}}(0|q), P_{\mathsf{U}|\mathsf{Q}}(1|q)\}.$$

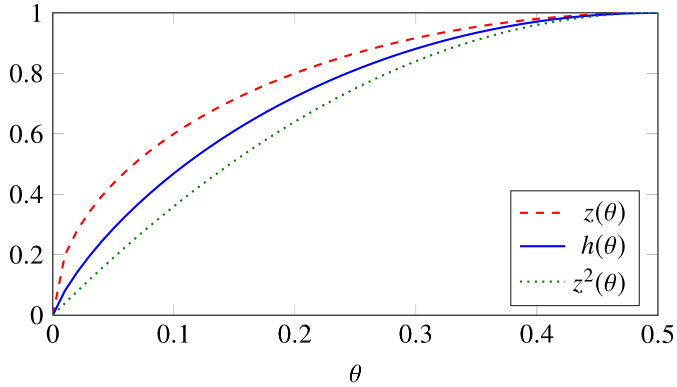

Fig. 3. Illustration that $z^2(\theta) \le h(\theta) \le z(\theta)$ for $0 \le \theta \le 1/2$.

Accordingly, $1 - \theta = \max\{P_{\mathsf{U}|\mathsf{Q}}(0|q), P_{\mathsf{U}|\mathsf{Q}}(1|q)\}$ and $\theta \in [0, 1/2]$. The various distribution parameters are expectations of functions of $\theta$:

$$\begin{aligned}\mathcal{P}_e(\mathsf{U}|\mathsf{Q}) &= \sum_q P_{\mathsf{Q}}(q)\theta,\\ \mathcal{K}(\mathsf{U}|\mathsf{Q}) &= \sum_q P_{\mathsf{Q}}(q)k(\theta),\\ \mathcal{H}(\mathsf{U}|\mathsf{Q}) &= \sum_q P_{\mathsf{Q}}(q)h(\theta),\\ \mathcal{Z}(\mathsf{U}|\mathsf{Q}) &= \sum_q P_{\mathsf{Q}}(q)z(\theta).\end{aligned}$$

We directly obtain the equality in (4a), as

$$\begin{aligned}\mathcal{K}(\mathsf{U}|\mathsf{Q}) &= \sum_q P_{\mathsf{Q}}(q)(1-2\theta)\\ &= 1 - 2\sum_q P_{\mathsf{Q}}(q)\theta\\ &= 1 - 2\mathcal{P}_e(\mathsf{U}|\mathsf{Q}).\end{aligned}$$

The inequality of (4a) is a consequence of Lemma 16, as

$$\begin{aligned}\mathcal{K}(\mathsf{U}|\mathsf{Q}) + \mathcal{H}(\mathsf{U}|\mathsf{Q}) &= \sum_q P_{\mathsf{Q}}(q)(k(\theta) + h(\theta))\\ &\ge 1.\end{aligned}$$

The right-most inequalities of (4b) and (4c) are immediate consequences of Lemma 15. Thus, we concentrate on the left-most inequalities.

For the left-most inequality of (4b), we employ Jensen's inequality for the convex-∪ function $x \mapsto x^2$ and the inequality $z^2(\theta) \le h(\theta)$ from Lemma 15 to obtain

$$\begin{aligned}\mathcal{Z}(\mathsf{U}|\mathsf{Q})^2 &= \left(\sum_q P_{\mathsf{Q}}(q)z(\theta)\right)^2\\ &\le \sum_q P_{\mathsf{Q}}(q)z^2(\theta)\\ &\le \sum_q P_{\mathsf{Q}}(q)h(\theta)\\ &= \mathcal{H}(\mathsf{U}|\mathsf{Q}).\end{aligned}$$

For the left-most inequality of (4c), observe that

$$\begin{aligned} z^2(\theta) + k^2(\theta) &= 4\theta(1-\theta) + (\theta - (1-\theta))^2 \\ &= \theta^2 + 2\theta(1-\theta) + (1-\theta)^2 \\ &= (\theta + (1-\theta))^2 \\ &= 1. \end{aligned}$$

Using Jensen's inequality twice for the convex-∪ function $x \mapsto x^2$,

$$\mathcal{Z}(\mathsf{U}|\mathsf{Q})^2 + \mathcal{K}(\mathsf{U}|\mathsf{Q})^2 \leq \sum_q P_{\mathsf{Q}}(q)(z^2(\theta) + k^2(\theta)) = 1.$$

This implies the left-most inequality of (4c). ■

*Proof of Lemma 2:* We obtain the joint distribution of $(\mathsf{U}, \mathsf{Q})$ by marginalizing $P_{\mathsf{U},\mathsf{Q},\mathsf{S}}$,

$$P_{\mathsf{U},\mathsf{Q}}(u, q) = \sum_s P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(u, q, s).$$

The triangle inequality yields (6a):

$$\begin{aligned} \mathcal{K}(\mathsf{U}|\mathsf{Q}) &= \sum_q |P_{\mathsf{U},\mathsf{Q}}(0, q) - P_{\mathsf{U},\mathsf{Q}}(1, q)| \\ &= \sum_q \left| \sum_s \left( P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(0, q, s) - P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(1, q, s) \right) \right| \\ &\leq \sum_{q,s} \left| P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(0, q, s) - P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(1, q, s) \right| \\ &= \mathcal{K}(\mathsf{U}|\mathsf{Q}, \mathsf{S}). \end{aligned}$$

We derive (6b) using the Cauchy-Schwartz inequality:

$$\begin{aligned} \mathcal{Z}(\mathsf{U}|\mathsf{Q}) &= 2 \sum_q \sqrt{P_{\mathsf{U},\mathsf{Q}}(0, q) P_{\mathsf{U},\mathsf{Q}}(1, q)} \\ &= 2 \sum_q \sqrt{\sum_s P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(0, q, s) \sum_{s'} P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(1, q, s')} \\ &\geq 2 \sum_{q,s} \sqrt{P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(0, q, s)} \sqrt{P_{\mathsf{U},\mathsf{Q},\mathsf{S}}(1, q, s)} \\ &= \mathcal{Z}(\mathsf{U}|\mathsf{Q}, \mathsf{S}). \end{aligned}$$

Inequality (6c) is a consequence of Jensen's inequality for the concave-∩ function $x \mapsto -x \log_2 x$. A proof can be found in [17, Theorem 2.6.5]. ■

## Appendix B
## Extension to the Non-Binary Case

Our results are readily extended to the non-binary case. Here, $\mathsf{X}_j$, $j \in \mathbb{Z}$ take values in an alphabet $\mathcal{U}$ with $|\mathcal{U}| = L$, where $L$ is prime.[9] As in [3], we use Arıkan's polarization transform in the non-binary case, replacing addition of $L$-ary numbers with modulo-$L$ addition. Thus (7) applies in the non-binary case; addition in (7a) and (7b) is modulo-$L$.

First, we extend the distribution parameters from Section III-A to non-binary $\mathsf{U}$. We do this while keeping key properties that allows their use in polar code analysis. Then, we consider their fast polarization. Fast polarization of the Bhattacharyya process was established in [4, Chapter 3]. We show that the total variation process satisfies the conditions for fast polarization required for Lemma 3.

### *A. Non-Binary Distribution Parameters*

The three distribution parameters we consider — Bhattacharyya parameter, total variation distance, and conditional entropy — were all defined for random variable pairs $(\mathsf{U}, \mathsf{Q})$ where $\mathsf{U}$ is binary. We now show how to extend them to the case where $\mathsf{U}$ may take values in an arbitrary finite alphabet $\mathcal{U}$. We denote $|\mathcal{U}| = L$.

There are two properties of the distribution parameters that are crucial for the analysis of polar codes. First, they are to take values in $[0, 1]$. Second, when each of them approaches one of the extreme values, so should the others. The suggested non-binary extension satisfies these properties. The extension for the Bhattacharyya parameter and conditional entropy are based on [3], which ensued the study of non-binary polar codes (see also [4, Chapter 3]).

Denote

$$\begin{aligned} \mathcal{Z}_L(\mathsf{U}|\mathsf{Q}) &= \sum_q \sum_{u' \neq u} \frac{P_{\mathsf{Q}}(q)}{L-1} \sqrt{P_{\mathsf{U}|\mathsf{Q}}(u|q) P_{\mathsf{U}|\mathsf{Q}}(u'|q)}, \\ \mathcal{K}_L(\mathsf{U}|\mathsf{Q}) &= \sum_q \sum_{u' \neq u} \frac{P_{\mathsf{Q}}(q)}{L-1} \frac{\left| P_{\mathsf{U}|\mathsf{Q}}(u|q) - P_{\mathsf{U}|\mathsf{Q}}(u'|q) \right|}{2}, \\ \mathcal{H}_L(\mathsf{U}|\mathsf{Q}) &= -\sum_q \sum_u P_{\mathsf{Q}}(q) P_{\mathsf{U}|\mathsf{Q}}(u|q) \log_L P_{\mathsf{U}|\mathsf{Q}}(u|q). \end{aligned}$$

As expected, when $L = 2$ these coincide with (1)–(3). All three parameters are in $[0, 1]$. This is well-known for the conditional entropy (see, e.g., [17, Chapter 2]); for the total variation distance and the Bhattacharyya parameter, see the proof of Lemma 17, below. The three parameters achieve their extreme values either when $P_{\mathsf{U}|\mathsf{Q}}(u|q) = 1/L$ for all $u$ or when there is some $u_0 \in \mathcal{U}$ such that $P_{\mathsf{U}|\mathsf{Q}}(u_0|q) = 1$ and $P_{\mathsf{U}|\mathsf{Q}}(u|q) = 0$ for $u \neq u_0$.

The consequences of Lemma 1 apply in the non-binary case as well. That is, when one of the three parameters approaches an extreme value, so do the other two. This is a consequence of the following lemma.

**Lemma 17.** *The non-binary total variation distance, probability of error, conditional entropy, and Bhattacharyya parameter are related by*

$$\mathcal{Z}_L(\mathsf{U}|\mathsf{Q})^2 \leq \mathcal{H}_L(\mathsf{U}|\mathsf{Q}) \leq \log_L(1 + (L-1)\mathcal{Z}_L(\mathsf{U}|\mathsf{Q})), \tag{40a}$$

$$1 - \mathcal{Z}_L(\mathsf{U}|\mathsf{Q}) \leq \mathcal{K}_L(\mathsf{U}|\mathsf{Q}) \leq \sqrt{1 - \mathcal{Z}_L(\mathsf{U}|\mathsf{Q})^2}. \tag{40b}$$

*Remark* 5. Inequality (40b) was also independently derived for the binary symmetric case in [20], using a different proof.

*Proof:* The inequalities in (40a) were derived in [4, Proposition 3.3]. Thus, we concentrate on showing (40b).

To see the right-most inequality of (40b), note that

$$\sum_q \sum_{u' \neq u} \frac{P_{\mathsf{Q}}(q)}{L(L-1)} = \sum_{q,u} \sum_{\underline{u'} \neq u} \frac{P_{\mathsf{Q}}(q)}{L(L-1)} = 1.$$

[9] When $L$ is not prime, one would need to consider other polarization kernels, which is beyond the scope of this work.

Thus, by Jensen's inequality,

$$\frac{\mathcal{Z}_L(\mathsf{U}|\mathsf{Q})^2}{L^2} \le \sum_{q,u}\sum_{u'\neq u} \frac{P_{\mathsf{Q}}(q)}{L(L-1)}\left(\sqrt{P_{\mathsf{U}|\mathsf{Q}}(u|q)P_{\mathsf{U}|\mathsf{Q}}(u'|q)}\right)^2,$$

$$\frac{\mathcal{K}_L(\mathsf{U}|\mathsf{Q})^2}{L^2} \le \sum_{q,u}\sum_{u'\neq u} \frac{P_{\mathsf{Q}}(q)}{L(L-1)}\left(\frac{P_{\mathsf{U}|\mathsf{Q}}(u|q)-P_{\mathsf{U}|\mathsf{Q}}(u'|q)}{2}\right)^2.$$

Next, observe that

$$\begin{aligned}\left(\sqrt{P_{\mathsf{U}|\mathsf{Q}}(u|q)P_{\mathsf{U}|\mathsf{Q}}(u'|q)}\right)^2 &+ \left(\frac{P_{\mathsf{U}|\mathsf{Q}}(u|q)-P_{\mathsf{U}|\mathsf{Q}}(u'|q)}{2}\right)^2\\ &= \left(\frac{P_{\mathsf{U}|\mathsf{Q}}(u|q)+P_{\mathsf{U}|\mathsf{Q}}(u'|q)}{2}\right)^2\end{aligned}$$

and that subject to the constraint $\sum_u P_{\mathsf{U}|\mathsf{Q}}(u|q) = 1$, we have

$$\sum_u\sum_{u'\neq u}\left(\frac{P_{\mathsf{U}|\mathsf{Q}}(u|q)+P_{\mathsf{U}|\mathsf{Q}}(u'|q)}{2}\right)^2 \le \frac{L(L-1)}{L^2}.$$

This can be seen using Lagrange multipliers; the maximum value is obtained with equality when $P_{\mathsf{U}|\mathsf{Q}}(u|q) = 1/L$ for all $u \in \mathcal{U}$. Thus, we obtain

$$\mathcal{Z}_L(\mathsf{U}|\mathsf{Q})^2 + \mathcal{K}_L(\mathsf{U}|\mathsf{Q})^2 \le 1,$$

which implies the right-most inequality of (40b). This also shows that indeed $\mathcal{Z}_L(\mathsf{U}|\mathsf{Q}) \le 1$ and $\mathcal{K}_L(\mathsf{U}|\mathsf{Q}) \le 1$.

For the left-most inequality of (40b), observe that for any $a, b \ge 0$ we have $\sqrt{ab} \ge \min\{a,b\}$, by which

$$\begin{aligned}\frac{|a-b|}{2} + \sqrt{ab} &= \frac{\max\{a,b\}-\min\{a,b\}}{2} + \sqrt{ab}\\ &\ge \frac{\max\{a,b\}-\min\{a,b\}+2\min\{a,b\}}{2}\\ &= \frac{\max\{a,b\}+\min\{a,b\}}{2}\\ &= \frac{a+b}{2}.\end{aligned}$$

Since

$$\sum_{u'\neq u} P_{\mathsf{U}|\mathsf{Q}}(u|q) = \sum_{u'\neq u} P_{\mathsf{U}|\mathsf{Q}}(u'|q) = L-1,$$

we have

$$\begin{aligned}\mathcal{Z}_L(\mathsf{U}|\mathsf{Q}) + \mathcal{K}_L(\mathsf{U}|\mathsf{Q}) &\ge \sum_q P_{\mathsf{Q}}(q)\sum_{u'\neq u}\frac{P_{\mathsf{U}|\mathsf{Q}}(u|q)+P_{\mathsf{U}|\mathsf{Q}}(u'|q)}{2(L-1)}\\ &= 1.\end{aligned}$$

This yields the left-most inequality of (40b). ■

Indeed, inequalities (40) imply that when either $\mathcal{Z}_L(\mathsf{U}|\mathsf{Q})$ or $\mathcal{H}_L(\mathsf{U}|\mathsf{Q})$ approach 0 or 1 then $\mathcal{K}_L(\mathsf{U}|\mathsf{Q})$ approaches 1 or 0, respectively, and vice versa.

In the binary case, the total-variation distance and the probability of error were related by (4a). In the non-binary case, the probability of error is given by

$$\mathcal{P}_{e,L}(\mathsf{U}|\mathsf{Q}) = \sum_q P_{\mathsf{Q}}(q)(1-\max_u P_{\mathsf{U}|\mathsf{Q}}(u|q)).$$

The non-binary probability of error and total variation distance are related, as shown in the following lemma.

**Lemma 18.** *The non-binary probability of error and total variation distance are related by*

$$\mathcal{K}_L(\mathsf{U}|\mathsf{Q}) \le 1 - \frac{2}{L-1}\mathcal{P}_{e,L}(\mathsf{U}|\mathsf{Q}).$$

*Proof:* Let $\mathcal{U} = \{0, 1, \dots, L-1\}$. Without loss of generality we assume that, for a given $q \in \mathcal{Q}$,

$$P_{\mathsf{U}|\mathsf{Q}}(0|q) \le P_{\mathsf{U}|\mathsf{Q}}(1|q) \le \cdots \le P_{\mathsf{U}|\mathsf{Q}}(L-1|q). \tag{41}$$

We then have

$$\begin{aligned}&\sum_{u'\neq u}\frac{|P_{\mathsf{U}|\mathsf{Q}}(u|q)-P_{\mathsf{U}|\mathsf{Q}}(u'|q)|}{2}\\ &\overset{(a)}{=} \sum_{u=0}^{L-1} uP_{\mathsf{U}|\mathsf{Q}}(u|q) - \sum_{u=0}^{L-1}(L-1-u)P_{\mathsf{U}|\mathsf{Q}}(u|q)\\ &\overset{(b)}{=} L - \sum_{u=0}^{L-1}(L-u)P_{\mathsf{U}|\mathsf{Q}}(u|q) - \sum_{u=0}^{L-1}(L-1-u)P_{\mathsf{U}|\mathsf{Q}}(u|q)\\ &= (L-1) - 2\sum_{u=0}^{L-1}(L-1-u)P_{\mathsf{U}|\mathsf{Q}}(u|q)\\ &\le (L-1) - 2\sum_{u=0}^{L-1}\min\{1, L-1-u\}P_{\mathsf{U}|\mathsf{Q}}(u|q)\\ &= (L-1) - 2\sum_{u=0}^{L-2}P_{\mathsf{U}|\mathsf{Q}}(u|q)\\ &\overset{(c)}{=} (L-1) - 2(1-\max_u P_{\mathsf{U}|\mathsf{Q}}(u|q)).\end{aligned}$$

To see (a), note that $|a-b| = \max\{a,b\} - \min\{a,b\}$. Using the ordering (41), we construct two $L \times L$ matrices: one with constant columns, with value $P_{\mathsf{U}|\mathsf{Q}}(u|q)$ in column $(u+1)$, and one with constant rows, with value $P_{\mathsf{U}|\mathsf{Q}}(u|q)$ in row $(u+1)$, $u = 0, 1, \dots, L-1$ . We compute the difference of the two matrices; the desired sum equals the sum of elements above the diagonal. Then, (b) is because $\sum uP_{\mathsf{U}|\mathsf{Q}}(u|q) + \sum(L-u)P_{\mathsf{U}|\mathsf{Q}}(u|q) = L$, and (c) is by the ordering (41) and since $\sum_u P_{\mathsf{U}|\mathsf{Q}}(u|q) = 1$. Thus, for any $q \in \mathcal{Q}$,

$$\sum_{u'\neq u}\frac{|P_{\mathsf{U}|\mathsf{Q}}(u|q)-P_{\mathsf{U}|\mathsf{Q}}(u'|q)|}{2(L-1)} \le 1 - \frac{2(1-\max_u P_{\mathsf{U}|\mathsf{Q}}(u|q))}{L-1}. \tag{42}$$

Using (42) in the definition of $\mathcal{K}_L(\mathsf{U}|\mathsf{Q})$ and recalling the expression for $\mathcal{P}_{e,L}(\mathsf{U}|\mathsf{Q})$, we obtain the desired inequality. ■

The following corollary tightens [4, Proposition 3.2].

**Corollary 19.** *The non-binary Bhattacharyya parameter upper-bounds the probability of error according to*

$$\mathcal{P}_{e,L}(\mathsf{U}|\mathsf{Q}) \le \frac{L-1}{2}\mathcal{Z}_L(\mathsf{U}|\mathsf{Q}).$$

*Proof:* This is a consequence of the left-hand inequality of (40b) and Lemma 18. ■

The non-binary distribution parameters are all natural extensions of their versions when $\mathsf{U}$ is binary. In particular, the non-binary parameters have the same form as their binary counterparts. As shown above, the consequences of Lemma 1 apply to the non-binary parameters as well. They also satisfy Lemma 2; the extension of its proof is straightforward. Thus, the non-binary distribution parameters may be used to define the relevant processes as in (8) and (25).

### *B. Polarization of the Distribution Parameters*

In the binary case, fast polarization is obtained by Lemma 3, which requires polarization bounds on the Bhattacharyya and total variation distance processes. In the non-binary case, the Bhattacharyya process and the total variation distance process are defined similarly to their binary counterparts, with the relevant parameters replaced with their non-binary form presented above. The relevant polarization bounds for the non-binary Bhattacharyya process were obtained in [4, Lemma 3.5]. We now establish polarization bounds for the total variation distance process that extend Proposition 4 to the non-binary case; we abuse notation and use $\mathsf{K}_n$ to denote the non-binary counterpart of the total variation distance process. Proposition 12 is similarly extended; we omit the derivation.

**Proposition 20.** *Assume that* $(\mathsf{X}_j, \mathsf{Y}_j)$*,* $j \in \mathbb{Z}$ *is a memoryless process, where* $\mathsf{X}_j \in \mathcal{U}$ *such that* $|\mathcal{U}| = L$*, and* $\mathsf{Y}_j \in \mathcal{Y}$*. Then,*

$$\mathsf{K}_{n+1} \leq \begin{cases} \dfrac{2(L-1)}{L}\mathsf{K}_n^2, & \textit{if } \mathsf{B}_{n+1} = 0, \\ \left(1 + \dfrac{L}{2}\right)\mathsf{K}_n, & \textit{if } \mathsf{B}_{n+1} = 1. \end{cases} \tag{43}$$

Observe that when $L = 2$, the right-hand-side of (43) coincides with that of (13).

*Proof:* As in Proposition 4, we fix $\mathsf{B}_1, \dots, \mathsf{B}_n$ and let $i - 1 = (\mathsf{B}_1\mathsf{B}_2 \cdots \mathsf{B}_n)_2$. This also fixes the value of $\mathsf{K}_n$. We denote $P_{\mathsf{U}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i}(u, v, q, r) = P(u, q)P(v, r)$. Slightly abusing notation, we further denote $P(u, q) = P(q)P(u|q)$. We set $\mathsf{T}_i = \mathsf{U}_i + \mathsf{V}_i$; this is modulo-$L$ addition, so

$$P_{\mathsf{T}_i,\mathsf{V}_i,\mathsf{Q}_i,\mathsf{R}_i}(t, v, q, r) = P(t - v, q)P(v, r),$$

where $t - v$ is computed modulo-$L$.

We shall need the following inequality:

$$\begin{aligned} &\sum_{u' \neq u} |P_{\mathsf{U}|\mathsf{Q}}(u|q) - P_{\mathsf{U}|\mathsf{Q}}(u'|q)| \\ &\quad \overset{(a)}{\geq} \left| \sum_{u' \neq u} (P_{\mathsf{U}|\mathsf{Q}}(u|q) - P_{\mathsf{U}|\mathsf{Q}}(u'|q)) \right| \\ &\quad \overset{(b)}{=} \left| (L-1)P_{\mathsf{U}|\mathsf{Q}}(u|q) - (1 - P_{\mathsf{U}|\mathsf{Q}}(u|q)) \right| \\ &\quad = L \left| P_{\mathsf{U}|\mathsf{Q}}(u|q) - \frac{1}{L} \right|. \end{aligned} \tag{44}$$

Here, (a) is by the triangle inequality and (b) is because $\sum_u P_{\mathsf{U}|\mathsf{Q}}(u|q) = 1$.

We compute $\mathsf{K}_{n+1}$ using (14). For the case $\mathsf{B}_{n+1} = 0$, note that

$$\begin{aligned} &\sum_{t' \neq t} \left| \sum_v P(v|r)(P(t - v|q) - P(t' - v|q)) \right| \\ &\quad \overset{(a)}{=} \sum_{t' \neq t} \left| \sum_v \left(P(v|r) - \frac{1}{L}\right)\left(P(t - v|q) - P(t' - v|q)\right) \right| \\ &\quad \overset{(b)}{\leq} \sum_{t' \neq t} \sum_v \left| \left(P(v|r) - \frac{1}{L}\right)\left(P(t - v|q) - P(t' - v|q)\right) \right| \\ &\quad \overset{(c)}{=} \sum_{t' \neq t} \sum_v \left| P(v|r) - \frac{1}{L} \right| \cdot \left| P(t - v|q) - P(t' - v|q) \right| \\ &\quad = \sum_v \left| P(v|r) - \frac{1}{L} \right| \cdot \sum_{t' \neq t} \left| P(t - v|q) - P(t' - v|q) \right| \\ &\quad \overset{(d)}{\leq} \frac{1}{L} \sum_{v' \neq v} \left| P(v|r) - P(v'|r) \right| \cdot \sum_{t' \neq t} \left| P(t|q) - P(t'|q) \right|, \end{aligned}$$

where (a) is because $\sum_v P(t - v|q) = \sum_v P(t' - v|q)$ for any $t, t'$, (b) is by the triangle inequality, (c) is because $|ab| = |a| \cdot |b|$, and (d) is by (44) and since the sum over $t, t'$ is unaffected by the shift in $v$. Thus,

$$\begin{aligned} &\mathcal{K}_L(\mathsf{T}_i|\mathsf{Q}_i, \mathsf{R}_i) \\ &\quad = \sum_{q,r} \frac{P(q)P(r)}{2(L-1)} \sum_{t \neq t'} \left| \sum_v P(v|r)(P(t - v|q) - P(t' - v|q)) \right| \\ &\quad \leq \frac{2(L-1)}{L}\mathsf{K}_n^2. \end{aligned}$$

Recalling (14), this proves the top inequality of (43).

For the case $\mathsf{B}_{n+1} = 1$, note that by (15) and the triangle inequality, when $v' \neq v$ we have

$$\begin{aligned} &2\left| P(t - v|q)P(v|r) - P(t - v'|q)P(v'|r) \right| \\ &\quad = \Big| \Big(P(t - v|q) + P(t - v'|q)\Big)\Big(P(v|r) - P(v'|r)\Big) \\ &\qquad + \Big(P(v|r) + P(v'|r)\Big)\Big(P(t - v|q) - P(t - v'|q)\Big) \Big| \\ &\quad \leq \Big(P(t - v|q) + P(t - v'|q)\Big) \cdot \left| P(v|r) - P(v'|r) \right| \\ &\qquad + \Big(P(v|r) + P(v'|r)\Big) \cdot \left| P(t - v|q) - P(t - v'|q) \right| \\ &\quad \leq \Big(P(t - v|q) + P(t - v'|q)\Big) \cdot \left| P(v|r) - P(v'|r) \right| \\ &\qquad + \left| P(t - v|q) - P(t - v'|q) \right|. \end{aligned}$$

The last inequality is due to the upper bound $P(v|r) + P(v'|r) \leq 1$ when $v' \neq v$. Hence,

$$\begin{aligned} &\sum_t \sum_{v' \neq v} |P(t - v|q)P(v|r) - P(t - v'|q)P(v'|r)| \\ &\quad \leq \sum_{v' \neq v} \left| P(v|r) - P(v'|r) \right| \\ &\qquad + \frac{1}{2} \sum_t \sum_{v' \neq v} \left| P(t - v|q) - P(t - v'|q) \right| \\ &\quad = \sum_{v' \neq v} \left| P(v|r) - P(v'|r) \right| + \frac{L}{2} \sum_{t' \neq t} \left| P(t|q) - P(t'|q) \right|. \end{aligned}$$

Consequently,

$$\begin{aligned} &\mathcal{K}_L(\mathsf{V}_i|\mathsf{T}_i, \mathsf{Q}_i, \mathsf{R}_i) \\ &\quad = \sum_{q,r,t} \sum_{v' \neq v} \frac{P(q)P(r)}{2(L-1)} |P(t - v|q)P(v|r) - P(t - v'|q)P(v'|r)| \\ &\quad \leq \left(1 + \frac{L}{2}\right)\mathsf{K}_n. \end{aligned}$$

This proves the bottom inequality of (43). ■

The bounds in (43) are of the form required in Lemma 3, allowing its use to establish fast polarization of the total variation distance process.

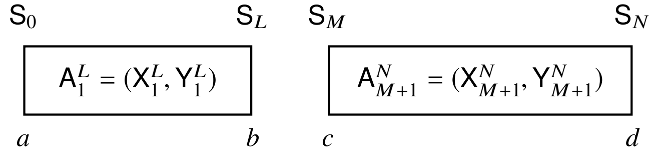
$\mathsf{S}_0$
$\mathsf{S}_L$
$\mathsf{S}_M$
$\mathsf{S}_N$
$\mathsf{A}_1^L = (\mathsf{X}_1^L, \mathsf{Y}_1^L)$
$\mathsf{A}_{M+1}^N = (\mathsf{X}_{M+1}^N, \mathsf{Y}_{M+1}^N)$
$a$
$b$
$c$
$d$

Fig. 4. Two blocks of a FAIM process, not necessarily of the same length. The initial state of the first block, $\mathsf{S}_0$, assumes value $a \in \mathcal{S}$. The final state of the first block, $\mathsf{S}_L$, assumes value $b \in \mathcal{S}$. The initial state of the second block, $\mathsf{S}_M$, assumes value $c \in \mathcal{S}$. The final state of the second block, $\mathsf{S}_N$, assumes value $d \in \mathcal{S}$.

## Appendix C
## Auxiliary Proofs for Section IV

We denote $\mathsf{A}_j = (\mathsf{X}_j, \mathsf{Y}_j)$, $j \in \mathbb{Z}$, with realization $\alpha_j$, and $\mathsf{A}_M^N = (\mathsf{X}_M^N, \mathsf{Y}_M^N)$ with realization $\alpha_M^N$. For brevity, we denote $P_{\mathsf{A}_M^N} \equiv P_{\mathsf{A}_M^N}(\alpha_M^N)$, and similarly $P_{\mathsf{S}_N} \equiv P_{\mathsf{S}_N}(s_N)$.

*Proof of Lemma 5:* The function $\psi(N)$ was defined in (19). We repeat the definition below using a notation that highlights the random variables at play. We deliberately do not use the notation (17), to explicitly show which random variables are being marginalized.

$$\psi(N) = \begin{cases} \max\limits_{a,b} \dfrac{P_{\mathsf{S}_N|\mathsf{S}_0}(b|a)}{P_{\mathsf{S}_0}(b)}, & \text{if } N > 0, \\[2ex] \max\limits_{a} \dfrac{1}{P_{\mathsf{S}_0}(a)}, & \text{if } N = 0. \end{cases}$$

Recall that by stationarity, $P_{\mathsf{S}_0} = P_{\mathsf{S}_N}$ for any $N$, so $P_{\mathsf{S}_N|\mathsf{S}_0}(b|a)/P_{\mathsf{S}_0}(b) = P_{\mathsf{S}_N|\mathsf{S}_0}(b|a)/P_{\mathsf{S}_N}(b)$.

Since $\mathsf{S}_j$, $j = 1, 2, \ldots$ is an aperiodic and irreducible stationary finite-state Markov chain, $\psi(N)$ is non-increasing and $\psi(N) \to 1$ as $N \to \infty$. This is evident from the properties of such Markov chains; for a formal proof of this statement, see [24, Theorem 7.14]. For such Markov chains $P_{\mathsf{S}_0}(a) > 0$ for any $a \in \mathcal{S}$, so $\psi(0) < \infty$.

It remains to show that $P_{\mathsf{A}_1^L, \mathsf{A}_{M+1}^N} \leq \psi(M-L) P_{\mathsf{A}_1^L} P_{\mathsf{A}_{M+1}^N}$. Consider first the case $M > L$. Denote by $a, b, c, d$ the *values* of states $\mathsf{S}_0, \mathsf{S}_L, \mathsf{S}_M$, and $\mathsf{S}_N$, respectively (see Figure 4). Then,

$$\begin{aligned}
&P_{\mathsf{A}_1^L, \mathsf{A}_{M+1}^N} \\
&\quad= \sum_{\alpha_{L+1}^M} P_{\mathsf{A}_1^L, \mathsf{A}_{L+1}^M, \mathsf{A}_{M+1}^N} \\
&\quad= \sum_{\alpha_{L+1}^M} \sum_{d,a} P_{\mathsf{A}_1^L, \mathsf{A}_{L+1}^M, \mathsf{A}_{M+1}^N, \mathsf{S}_N | \mathsf{S}_0} P_{\mathsf{S}_0} \\
&\quad= \sum_{\substack{d,c,\\ b,a}} \sum_{\alpha_{L+1}^M} P_{\mathsf{A}_{M+1}^N, \mathsf{S}_N | \mathsf{S}_M} P_{\mathsf{A}_{L+1}^M, \mathsf{S}_M | \mathsf{S}_L} P_{\mathsf{A}_1^L, \mathsf{S}_L | \mathsf{S}_0} P_{\mathsf{S}_0} \\
&\quad= \sum_{\substack{d,c,\\ b,a}} P_{\mathsf{A}_{M+1}^N, \mathsf{S}_N | \mathsf{S}_M} \left( \sum_{\alpha_{L+1}^M} P_{\mathsf{A}_{L+1}^M, \mathsf{S}_M | \mathsf{S}_L} \right) P_{\mathsf{A}_1^L, \mathsf{S}_L | \mathsf{S}_0} P_{\mathsf{S}_0} \\
&\quad= \sum_{\substack{d,c,\\ b,a}} P_{\mathsf{A}_{M+1}^N, \mathsf{S}_N | \mathsf{S}_M} P_{\mathsf{S}_M | \mathsf{S}_L} P_{\mathsf{A}_1^L, \mathsf{S}_L | \mathsf{S}_0} P_{\mathsf{S}_0} \\
&\quad= \sum_{\substack{d,c,\\ b,a}} P_{\mathsf{A}_{M+1}^N, \mathsf{S}_N | \mathsf{S}_M} P_{\mathsf{S}_M} \frac{P_{\mathsf{S}_M | \mathsf{S}_L}}{P_{\mathsf{S}_M}} P_{\mathsf{A}_1^L, \mathsf{S}_L | \mathsf{S}_0} P_{\mathsf{S}_0} \\
&\quad\leq \psi(M-L) \sum_{\substack{d,c,\\ b,a}} P_{\mathsf{A}_{M+1}^N, \mathsf{S}_N | \mathsf{S}_M} P_{\mathsf{S}_M} P_{\mathsf{A}_1^L, \mathsf{S}_L | \mathsf{S}_0} P_{\mathsf{S}_0} \\
&\quad= \psi(M-L) P_{\mathsf{A}_1^L} P_{\mathsf{A}_{M+1}^N}.
\end{aligned}$$

We proceed similarly for the case $M = L$. Again, $a$ and $d$ represent the *values* of states $\mathsf{S}_0$ and $\mathsf{S}_N$. Both $b$ and $b'$ represent values of state $\mathsf{S}_L$; this distinction is to distinguish the summation variables of two different sums over values of $\mathsf{S}_L$. Thus,

$$\begin{aligned}
P_{\mathsf{A}_1^L, \mathsf{A}_{L+1}^N} &= \sum_{\substack{a,b,\\ d}} P_{\mathsf{A}_{L+1}^N, \mathsf{S}_N | \mathsf{S}_L} \frac{P_{\mathsf{S}_L}}{P_{\mathsf{S}_L}} P_{\mathsf{A}_1^L, \mathsf{S}_L | \mathsf{S}_0} P_{\mathsf{S}_0} \\
&\leq \psi(0) \sum_{d,b} P_{\mathsf{A}_{L+1}^N, \mathsf{S}_N | \mathsf{S}_L} P_{\mathsf{S}_L} \cdot \left( \sum_{b',a} P_{\mathsf{A}_1^L, \mathsf{S}_L | \mathsf{S}_0} P_{\mathsf{S}_0} \right) \\
&= \psi(0) P_{\mathsf{A}_1^L} P_{\mathsf{A}_{L+1}^N};
\end{aligned}$$

where the inequality is because $P_{\mathsf{A}_1^L, \mathsf{S}_L | \mathsf{S}_0} \leq \sum_{b'} P_{\mathsf{A}_1^L, \mathsf{S}_L | \mathsf{S}_0}$. ■

*Proof of Lemma 6:* Due to aperiodicity and irreducibility of the state sequence, $P_{\mathsf{S}_M}(a) > 0$ for any $a \in \mathcal{S}$. By the Markov property,

$$\begin{aligned}
P_{\mathsf{A}_1^M, \mathsf{A}_{M+1}^N | \mathsf{S}_M} &= \frac{P_{\mathsf{S}_M, \mathsf{A}_1^M, \mathsf{A}_{M+1}^N}}{P_{\mathsf{S}_M}} \\
&= \frac{P_{\mathsf{S}_M} \cdot P_{\mathsf{A}_1^M | \mathsf{S}_M} \cdot P_{\mathsf{A}_{M+1}^N | \mathsf{S}_M, \mathsf{A}_1^M}}{P_{\mathsf{S}_M}} \\
&= P_{\mathsf{A}_1^M | \mathsf{S}_M} \cdot P_{\mathsf{A}_{M+1}^N | \mathsf{S}_M}.
\end{aligned}$$

This proves (20a).

To derive (20b), some more care is required to avoid division by 0. By the Markov property,

$$\begin{aligned}
&P_{\mathsf{S}_0, \mathsf{S}_M, \mathsf{S}_N} \cdot P_{\mathsf{A}_1^M, \mathsf{A}_{M+1}^N | \mathsf{S}_0, \mathsf{S}_M, \mathsf{S}_N} \\
&\quad= P_{\mathsf{S}_0, \mathsf{S}_M, \mathsf{S}_N, \mathsf{A}_1^M, \mathsf{A}_{M+1}^N} \\
&\quad= P_{\mathsf{S}_0, \mathsf{S}_M} \cdot P_{\mathsf{A}_1^M | \mathsf{S}_0, \mathsf{S}_M} \cdot P_{\mathsf{S}_N, \mathsf{A}_{M+1}^N | \mathsf{S}_0, \mathsf{S}_M, \mathsf{A}_1^M} \\
&\quad= P_{\mathsf{S}_0, \mathsf{S}_M} \cdot P_{\mathsf{A}_1^M | \mathsf{S}_0, \mathsf{S}_M} \cdot P_{\mathsf{S}_N, \mathsf{A}_{M+1}^N | \mathsf{S}_M} \\
&\quad= P_{\mathsf{S}_0, \mathsf{S}_M} \cdot P_{\mathsf{A}_1^M | \mathsf{S}_0, \mathsf{S}_M} \cdot P_{\mathsf{S}_N | \mathsf{S}_M} \cdot P_{\mathsf{A}_{M+1}^N | \mathsf{S}_M, \mathsf{S}_N} \\
&\quad= P_{\mathsf{S}_0, \mathsf{S}_M} \cdot P_{\mathsf{S}_N | \mathsf{S}_M, \mathsf{S}_0} \cdot P_{\mathsf{A}_1^M | \mathsf{S}_0, \mathsf{S}_M} \cdot P_{\mathsf{A}_{M+1}^N | \mathsf{S}_M, \mathsf{S}_N} \\
&\quad= P_{\mathsf{S}_0, \mathsf{S}_M, \mathsf{S}_N} \cdot P_{\mathsf{A}_1^M | \mathsf{S}_0, \mathsf{S}_M} \cdot P_{\mathsf{A}_{M+1}^N | \mathsf{S}_M, \mathsf{S}_N}.
\end{aligned}$$

Recalling the definition of conditional probability [22, Section 33], this implies (20b). ■

## References

[1] 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.

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

[3] E. Şaşoğlu, E. Telatar, and E. Arıkan, "Polarization for arbitrary discrete memoryless channels," in *2009 IEEE Information Theory Workshop*, October 2009, pp. 144–148.

[4] E. Şaşoğlu, "Polar Coding Theorems for Discrete Systems," Ph.D. dissertation, School Comput. Commun. Sci., EPFL, Lausanne, Switzerland, 2011.

[5] S. B. Korada and R. L. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1751–1768, April 2010.

[6] E. Arıkan, "Source polarization," in *2010 IEEE Int. Sym. on Information Theory*, June 2010, pp. 899–903.

[7] E. Hof and S. Shamai, "Secrecy-achieving polar-coding," in *2010 IEEE Information Theory Workshop*, August 2010, pp. 1–5.

[8] H. Mahdavifar and A. Vardy, "Achieving the secrecy capacity of wiretap channels using polar codes," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6428–6443, October 2011.

[9] 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.

[10] E. Arıkan, N. ul Hassan, M. Lentmaier, G. Montorsi, and J. Sayir, "Challenges and some new directions in channel coding," *Journal of Communications and Networks*, vol. 17, no. 4, pp. 328–338, August 2015.

[11] E. Şaşoğlu and I. Tal, "Polar coding for processes with memory," in *2016 IEEE Int. Sym. on Information Theory (ISIT)*. IEEE, 2016, pp. 225–229.

[12] E. Şaşoğlu and I. Tal, "Polar coding for processes with memory," 2016. [Online]. Available: http://arxiv.org/abs/1602.01870

[13] 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.

[14] 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

[15] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, Inc., 1968.

[16] B. Marcus, R. Roth, and P. Siegel, "Constrained systems and coding for recording channels," in *Handbook of Coding Theory*, V. Pless and W. Huffman, Eds. Amsterdam: Elsevier, 1998, pp. 1635–1764.

[17] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[18] I. Tal, "A simple proof of fast polarization," *IEEE Transactions on Information Theory*, vol. 63, no. 12, pp. 7617–7619, Dec 2017.

[19] M. Alsan, "Re-proving Channel Polarization Theorems," Ph.D. dissertation, IC, Lausanne, 2015. [Online]. Available: https://infoscience.epfl.ch/record/203886/files/EPFL_TH6403.pdf

[20] I. Dumer, "Polar codes with a stepped boundary," in *2017 IEEE Int. Sym. on Information Theory (ISIT)*, June 2017, pp. 2618–2622.

[21] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, EPFL, Lausanne, 2009. [Online]. Available: https://infoscience.epfl.ch/record/138655/files/EPFL_TH4461.pdf

[22] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1995.

[23] 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.

[24] R. Bradley, *Introduction to Strong Mixing Conditions*. Kendrick Press, 2007, vol. 1.

[25] A. Beck, *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. PA, USA: SIAM, 2014.

**Boaz Shuval** (M'09–S'17) was born in Haifa, Israel, in 1980. He received the B.Sc. and M.Sc. degrees in electrical engineering in 2002 and 2011, respectively, both from Technion — Israel Institute of Technology, Israel, where he is currently pursuing the PhD degree in electrical engineering. During 2002–2009 he worked as an engineer in the Israeli Defense Forces. Since 2009 he has been working as a researcher at Rafael — Advanced Defense Systems, Israel.

**Ido Tal** (S'05–M'08–SM'18) was born in Haifa, Israel, in 1975. He received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Technion — Israel Institute of Technology, Haifa, Israel, in 1998, 2003 and 2009, respectively. During 2010–2012 he was a postdoctoral scholar at the University of California at San Diego. In 2012 he joined the Electrical Engineering Department at Technion. His research interests include constrained coding and error-control coding. He received the IEEE Joint Communications Society/Information Theory Society Paper Award (jointly with Alexander Vardy) for the year 2017.