Polar Coding for Processes with Memory

Eren Şaşoğlu
Intel Corporation
Santa Clara, CA, USA
eren.sasoglu@intel.com

Ido Tal
Department of Electrical Engineering
Techion, Haifa 32000, Israel
idotal@ee.technion.ac.il

***Abstract*—We study polar coding over channels and sources with memory. We show that ψ-mixing processes polarize under the standard transform, and that the rate of polarization to deterministic distributions is roughly $O(2^{-\sqrt{N}})$ as in the memoryless case, where N is the blocklength. This implies that the error probability guarantees of polar channel and source codes extend to a large class of models with memory, including finite-order Markov sources and finite-state channels.**

***Index Terms*—Channels with memory, polar codes, periodic processes, strong polarization.**

I. Introduction

Polar codes were invented by Arıkan [1] as a low-complexity method to achieve the capacity of symmetric binary-input memoryless channels. The technique that underlies these codes, called *polarization*, is quite versatile, and has since been applied to numerous classical memoryless problems in information theory.

Many practical sources and channels are not well-described by memoryless models. In wireless communication, for example, memory in the form of intersymbol interference is quite prominent due to multipath propagation. In practice, this type of memory is most commonly handled by eliminating it, by augmenting the transmitter/receiver appropriately to create an overall memoryless channel. Memoryless coding techniques are then used for communication. Channel equalization and OFDM techniques are perhaps the most notable examples of this approach.

In contrast, here we are interested in whether polar coding can be used *directly* on channels and sources with memory, which may help simplify system design. In polarization theory, little is known for such settings. In particular, it was shown in [2, Chapter 5] that the standard transform polarizes strongly mixing processes with finite memory. In [3], it was shown that the successive cancellation decoding complexity of polar codes scales with the number of states of the underlying process, and thus is practical if the amount of memory in the system is modest. Whether polarization takes place sufficiently fast to yield a coding theorem has been left open, however.

Here, we first give a simpler proof of polarization than the one given in [2], for the more general class of ψ-mixing processes. We then show that the asymptotic rate of polarization to deterministic distributions is as in the memoryless case. This lets us conclude that the usual error probability guarantees of polar channel and source codes carry over to processes with memory, including well-behaved Markov sources as well as finite-state channels. For example, the results here imply that polar codes achieve the capacity of the Gilbert–Elliot channel (see [4], [5], and also [6]).

This work was done when Eren Şaşoğlu was at the Technion in June–July 2015.

II. Setting

Let (X_i, Y_i, S_i), $i \in \mathbb{Z}$, be a stationary process, where Y_i and S_i take values in finite alphabets $\mathcal{Y}$ and $\mathcal{S}$. We assume $X_i \in \{0, 1\}$ in order to keep the notation simple, but the results here can be generalized to arbitrary finite alphabets using standard techniques. See, for example, [2, Chapter 3].

We think of X_i as a sequence to be estimated, and Y_i as a sequence of observations related to X_i. In particular, X_i may be the input sequence to a communication channel, and Y_i the corresponding output. Alternatively, X_i may be the output of a data source to be compressed, and Y_i the side information available to the decompressor. A (possibly hidden) state sequence S_i may underlie the channel or the source. Frequently, one assumes that the pair (X_i, Y_i) is independent of the history $(X_1^{i-1}, Y_1^{i-1}, S_1^{i-1})$ conditioned on the present state S_i.

We assume throughout that the process (X_i, Y_i, S_i) is ψ-mixing. We follow[1] [7, Page 169] and say that a process T_i is ψ-mixing if there exists a sequence $\psi_k \to 1$ as $k \to \infty$ such that

$$\Pr(A \cap B) \leq \psi_k \Pr(A) \Pr(B) \tag{1}$$

for all $A \in \sigma(T_{-\infty}^0)$ and $B \in \sigma(T_{k+1}^\infty)$, where $\sigma(\cdot)$ denotes the sigma-field generated by its argument. Therefore, ψ-mixing implies that all pairs of events that are sufficiently far apart are almost independent. Note that the dependence of ψ_k on events A and B is only through the distance k between them.

Many source and channel models of practical importance are captured by ψ-mixing. In particular,

(i) an independent and identically distributed (i.i.d.) source X_i is ψ-mixing.
(ii) A finite-order, stationary, irreducible, aperiodic Markov source X_i is ψ-mixing.
(iii) Let X_i be a stationary source with state S_i, where the next source symbol and state depend only on their current values. That is,

$$p(s_{i+1}, x_i | s_{-\infty}^i, x_{-\infty}^{i-1}) = p(s_{i+1}, x_i | s_i, x_{i-1}) \ .$$

[1] To the best of our understanding, the first displayed equation on page 169 of [7] should be "$\sum_v \mu(uvw) \leq \cdots$".

The process (S_i, X_i) is Markov, and therefore if it is also irreducible and aperiodic, then it is ψ-mixing by (ii), and therefore so is X_i. This model covers sources generated by a hidden Markov state sequence, described by the conditional distributions

$$p(s_i, x_i | s_{-\infty}^{i-1}, x_{-\infty}^{i-1}) = p(s_i|s_{i-1}) p(x_i|s_i) \ .$$

(iv) If X_i is an i.i.d. input sequence to a discrete memoryless channel and Y_i is the output sequence, then (X_i, Y_i) is i.i.d. and therefore ψ-mixing by (i).

(v) Let W be a finite-state channel with input sequence X_i, output sequence Y_i, and state sequence S_i [8], all taking values in finite but otherwise arbitrary sets. The current output and the next state of the channel depend only on the current state and input:

$$p(s_i, y_i | x_{-\infty}^{i-1}, s_{-\infty}^{i-1}, y_{-\infty}^{i-1}) = W(s_i, y_i | x_{i-1}, s_{i-1}).$$

If the input X_i is Markov, then so is the process (X_i, Y_i, S_i), and thus it is also ψ-mixing.

The parameter ψ_0 plays an important role in this paper, and can be computed easily for all of the cases above [7, Page 169].

We are interested in the effects of the standard polar transform on the process. For this purpose, define the conditional entropy rate of X_i as

$$\begin{aligned}\mathcal{H}_{X|Y} &= \lim_{N\to\infty} \frac{1}{N} H(X_1^N | Y_1^N) \\ &= \lim_{N\to\infty} \frac{1}{N} H(X_1^N, Y_1^N) - \lim_{N\to\infty} \frac{1}{N} H(Y_1^N)\end{aligned}$$

The right-hand-side limits exist due to stationarity [9, Theorem 4.2.1]. We let $U_1^N = X_1^N \mathsf{B}_N \mathsf{G}_N$, where $N = 2^n$, $n = 1, 2, \ldots$, G_N is the nth Kronecker power of $\left(\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right)$ and B_N is the $N \times N$ bit-reversal matrix. We also define

$$Z(A|B) = 2 \sum_{b \in \mathcal{B}} \sqrt{p_{AB}(0, b) p_{AB}(1, b)}$$

for arbitrary random variables $A \in \{0, 1\}$ and B. It is well known that $Z(A|B)$, sometimes called the Bhattacharyya parameter, upper bounds the error probability of optimally guessing A by observing B. See, for example [2, Proposition 2.2].

The main results of this paper are the following.

Theorem 1 (Polarization). *If $\psi_0 < \infty$, then for all $\epsilon > 0$*

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N} \big|\{i : H(U_i | U_1^{i-1} Y_1^N) > 1 - \epsilon\}\big| &= \mathcal{H}_{X|Y}, \\ \lim_{N\to\infty} \frac{1}{N} \big|\{i : H(U_i | U_1^{i-1} Y_1^N) < \epsilon\}\big| &= 1 - \mathcal{H}_{X|Y}.\end{aligned}$$

Theorem 2 (Fast polarization). *If $\psi_0 < \infty$, then for all $\beta < 1/2$*

$$\lim_{N\to\infty} \frac{1}{N} \big|\{i : Z(U_i | U_1^{i-1} Y_1^N) < 2^{-N^\beta}\}\big| = 1 - \mathcal{H}_{X|Y}.$$

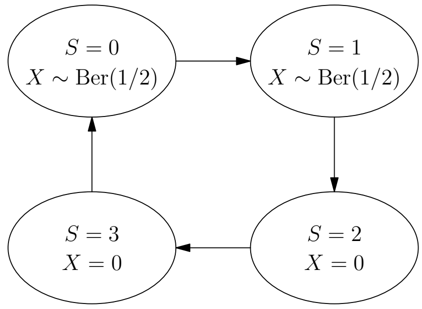

Fig. 1. Non polarizing process with period 4. Output is Bernoulli 1/2 for two phases and identically zero for next two phases.

Theorem 3 (Periodic processes may not polarize)**.** *The stationary periodic Markov process depicted in Figure 1 does not polarize. Indeed, for all $\frac{5N}{8} < i \le \frac{6N}{8}$,*

$$\left| H(U_i | U_1^{i-1}) - \frac{1}{2} \right| \le \epsilon_N \ , \quad \lim_{N\to\infty} \epsilon_N = 0 \ . \tag{2}$$

We will prove these claims in the following sections. Throughout, we will use the shorthand

$$\begin{aligned}H^{\mathbf{b}} &= H(U_i | Y_1^N U_1^{i-1}), \\ Z^{\mathbf{b}} &= Z(U_i | Y_1^N U_1^{i-1}),\end{aligned}$$

where $b \in \{0, 1\}^n$ is the n-bit binary expansion of $i - 1 \in \{0, \ldots, N - 1\}$. We will omit the range of indices when it is clear from context. The following are immediate from the definition of $\mathsf{B}_N \mathsf{G}_N$:

$$\begin{aligned}H^{\mathbf{b}0} &= H(U_{2i-1} | Y_1^{2N} U_1^{2i-2}) \\ H^{\mathbf{b}1} &= H(U_{2i} | Y_1^{2N} U_1^{2i-1})\end{aligned}$$

for all $\mathbf{b} \in \{0, 1\}^n$. Of course, the above also holds when the H are replaced by Z's. Further, for all lengths n, one can induce the uniform distribution on the set of $H^{\mathbf{b}}$'s and $Z^{\mathbf{b}}$'s by taking a sequence $B_1, B_2, \ldots$ of i.i.d. $\mathrm{Ber}(1/2)$ random variables and considering the random variables $H_n = H^{B_1 \ldots B_n}$ and $Z_n = Z^{B_1 \ldots B_n}$. Theorems 1 and 2 are then equivalent to

Theorem 4. *If $\psi_0 < \infty$, then for all $\epsilon > 0$*

$$\begin{aligned}\lim_{n\to\infty} P(H_n > 1 - \epsilon) &= \mathcal{H}_{X|Y}, \\ \lim_{n\to\infty} P(H_n < \epsilon) &= 1 - \mathcal{H}_{X|Y}.\end{aligned}$$

Theorem 5. *If $\psi_0 < \infty$, then for all $\beta < 1/2$*

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

III. Proof of Theorem 1

We will use the following shorthand in the rest of the paper:

$$\begin{aligned}U_1^N &= X_1^N \mathsf{B}_N \mathsf{G}_N \\ V_1^N &= X_{N+1}^{2N} \mathsf{B}_N \mathsf{G}_N \\ Q_i &= Y_1^N U_1^{i-1} \\ R_i &= Y_{N+1}^{2N} V_1^{i-1}\end{aligned} \tag{3}$$

For the proof, we take the somewhat standard approach of showing that H_n converges almost surely to a $\{0,1\}$-valued random variable. Recall that we have defined H_n through

$$H_n = H(U_i|Y_1^N U_1^{i-1}) \qquad \text{whenever } (B_1 \dots B_n)_2 = i-1.$$

Observe that for a given realization of H_n, we have

$$H_{n+1} = \begin{cases} H(U_i+V_i|Q_i,R_i) & \text{if } B_{n+1}=0 \\ H(V_i|Q_i,R_i,U_i+V_i) & \text{if } B_{n+1}=1 \end{cases}.$$

Further, since we have

$$\begin{aligned} H(U_i+V_i|Q_i,R_i) &+ H(V_i|Q_i,R_i,U_i+V_i) \\ &= H(U_i,V_i|Q_i,R_i) \le 2H(U_i|Q_i) \end{aligned}$$

and since $H_n \in [0,1]$, it follows that $H_1, H_2, \dots$ is a bounded supermartingale and thus converges almost surely to a $[0,1]$-valued random variable H_∞. It therefore remains to show that $H_\infty \in \{0,1\}$ almost surely. For this purpose, we will show that for all $\xi > 0$ there exists $\gamma(\xi) > 0$ such that

$$\begin{gathered} H(U_i|Q_i) \in (2\xi, 1-2\xi) \\ \text{implies} \\ H(U_i+V_i|Q_i,R_i) - H(U_i|Q_i) > \gamma(\xi) , \end{gathered} \tag{4}$$

for almost all i. That is, for a fraction of $i \in \{1,\dots,N\}$ approaching 1 as $N \to \infty$.

The theorem will follow from this claim, since (4) is equivalent to saying that if H_n is bounded away from 0 and 1, then $H_{n+1} - H_n$ is almost surely bounded away from 0. Therefore since H_n converges almost surely, it can do so only to 0 or 1.

We now show (4). We know from [2, Chapter 3] that the claim would hold for all N and i if (X_1^N, Y_1^N) and $(X_{N+1}^{2N}, Y_{N+1}^{2N})$ were independent. Our purpose here is to show that in the present setting there is sufficient independence between various random variables in neighboring blocks to imply (4). (This is essentially the approach taken in [2, Chapter 5], although the proof here is simpler and more general.) In particular, we will need the following independence results.

Lemma 6. *If $\psi_0 < \infty$, then for any $\epsilon > 0$, the fraction of indices i for which*

$$\begin{gathered} I(U_i;R_i|Q_i) < \epsilon \\ I(V_i;Q_i|R_i) < \epsilon \\ I(U_i;V_i|Q_i,R_i) < \epsilon \end{gathered}$$

approaches 1 as $N \to \infty$.

Proof. We only prove the first and the third inequality, the second follows by symmetry. We have

$$\begin{aligned} \log(\psi_0) &\ge E\left[\log \frac{p_{X_1^{2N}Y_1^{2N}}}{p_{X_1^N Y_1^N} \cdot p_{X_{N+1}^{2N}Y_{N+1}^{2N}}}\right] \\ &= I(X_1^N Y_1^N; X_{N+1}^{2N} Y_{N+1}^{2N}) \\ &\ge I(U_1^N; V_1^N Y_{N+1}^{2N} | Y_1^N) \\ &= \sum_{i=1}^N I(U_i; V_1^N Y_{N+1}^{2N} | Y_1^N U_1^{i-1}). \end{aligned}$$

Since all terms inside the sum are non-negative, it follows that at most $\sqrt{\log(\psi_0)N}$ (a vanishing fraction) of them are at most $\sqrt{\log(\psi_0)/N}$. Observing that the ith term is greater than both $I(U_i;R_i|Q_i)$ and $I(U_i;V_i|Q_i,R_i)$ concludes the proof. $\square$

Lemma 7. *Let (X_i, Y_i) be stationary and ψ-mixing. For all $\xi > 0$, there exists N_0 and $\delta(\xi) > 0$ such that for all $N > N_0$ and all $\{0,1\}$-valued random variables $A = f(X_1^N, Y_1^N)$ and $B = f(X_{N+1}^{2N}, Y_{N+1}^{2N})$*

$$p_A(0) \in (\xi, 1-\xi) \quad \textit{implies} \quad p_{AB}(0,1) > \delta(\xi).$$

Proof. Define $C = f(X_{2N+1}^{3N}, Y_{2N+1}^{3N})$. We have

$$\begin{aligned} 2p_{AB}(0,1) &= p_{AB}(0,1) + p_{BC}(0,1) \\ &\ge p_{ABC}(0,1,1) + p_{ABC}(0,0,1) \\ &= p_{AC}(0,1) \\ &= p_A(0) - p_{AC}(0,0) \\ &\ge p_A(0)(1 - \psi_N p_C(0)) \\ &= p_A(0)(1 - \psi_N p_A(0)) \end{aligned}$$

where the first and last equalities are due to stationarity. Since $p_A(0) \in (\xi, 1-\xi)$ and $\psi_N \to 1$, it follows that there exists N_0 such that the last term is away from 0 for all $N > N_0$. Further, since ψ_N is independent of A, so is N_0. This yields the claim. $\square$

We are ready to complete the proof by showing (4). Observe that we need to show the claim for arbitrarily small *but fixed* ξ. Let $(\tilde{U}_i, \tilde{V}_i)$ be random variables with

$$\begin{aligned} p_{\tilde{U}_i\tilde{V}_iQ_iR_i}&(u_i,v_i,q_i,r_i) \\ &= p_{U_i|Q_i}(u_i|q_i) p_{V_i|R_i}(v_i|r_i) p_{Q_iR_i}(q_i,r_i) . \end{aligned}$$

By definition,

$$H(\tilde{U}_i|Q_iR_i) = H(\tilde{U}_i|Q_i) = H(U_i|Q_i) .$$

A corollary of Lemma 6 is

Corollary 8. *If $\psi_0 < \infty$, then any $\epsilon > 0$, the fraction of indices i for which*

$$|H(\tilde{U}_i + \tilde{V}_i|Q_iR_i) - H(U_i+V_i|Q_i,R_i)| < \epsilon$$

approaches 1 as $N \to \infty$.

Therefore, it suffices to show that

$$\begin{gathered} H(\tilde{U}_i|Q_iR_i) \in (2\xi, 1-2\xi) \\ \text{implies} \\ H(\tilde{U}_i+\tilde{V}_i|Q_iR_i) - H(\tilde{U}_i|Q_iR_i) > 2\gamma(\xi) \end{gathered} \tag{5}$$

for all i in order to complete the proof. In order to do so, we will use the following fact, whose proof follows from convexity of binary entropy and is thus is omitted.

Lemma 9. *Let A and B be independent binary random variables. For every $\xi > 0$, there exists $\Delta(\xi) > 0$ such that*

$$\begin{gathered} \max\{H(A), H(B)\} > \xi \quad \textit{and} \\ \min\{H(A), H(B)\} < 1 - \xi \end{gathered}$$

imply

$$H(A+B) > \frac{H(A)+H(B)}{2} + \Delta(\xi).$$

For a given i, define the random variables $H_{Q_i}(\tilde{U}_i)$, $H_{R_i}(\tilde{V}_i)$, and $H_{Q_iR_i}(\tilde{U}_i+\tilde{V}_i)$ that take the values

$$\begin{aligned} H_{Q_i}(\tilde{U}_i) &= H(\tilde{U}_i|Q_i=q_i) \\ H_{R_i}(\tilde{V}_i) &= H(\tilde{V}_i|R_i=r_i) \\ H_{Q_iR_i}(\tilde{U}_i+\tilde{V}_i) &= H(\tilde{U}_i+\tilde{V}_i|(Q_i,R_i)=(q_i,r_i)) \end{aligned}$$

whenever

$$(Q_i,R_i)=(q_i,r_i).$$

Note that (5) is equivalent to:

$$\begin{aligned} &E[H_{Q_i}(\tilde{U}_i)] \in (2\xi, 1-2\xi) \text{ implies} \\ &\qquad E[H_{Q_iR_i}(\tilde{U}_i+\tilde{V}_i) - H_{Q_i}(\tilde{U}_i)] \geq 2\gamma(\xi). \end{aligned}$$

We take $2\gamma(\xi) = \delta(\xi)\Delta(\xi)$, where $\delta(\xi)$ and $\Delta(\xi)$ are as in Lemmas 7 and 9, respectively. We will be done if we can show that $E[H_{Q_i}(\tilde{U}_i)] \in (2\xi, 1-2\xi)$ implies

$$\begin{aligned} P\Big(&\max\{H_{Q_i}(\tilde{U}_i), H_{R_i}(\tilde{V}_i)\} > \xi \text{ and} \\ &\min\{H_{Q_i}(\tilde{U}_i), H_{R_i}(\tilde{V}_i)\} < 1-\xi \ \Big) > \delta(\xi)\ . \end{aligned} \tag{6}$$

Indeed, Lemma 9, and stationarity imply that

$$\begin{aligned} &E[H_{Q_iR_i}(\tilde{U}_i+\tilde{V}_i) - H_{Q_i}(\tilde{U}_i)] \\ &\quad = E\left[H_{Q_iR_i}(\tilde{U}_i+\tilde{V}_i) - \frac{H_{Q_i}(\tilde{U}_i)+H_{R_i}(\tilde{V}_i)}{2}\right] \\ &\qquad\qquad\qquad \geq \delta(\xi)\Delta(\xi)\ . \end{aligned}$$

Let us assume without loss of generality that $\delta(\xi) < \xi$. Thus, if $P(H_{Q_i}(\tilde{U}_i) \in (\xi, 1-\xi)) \geq \xi$, then (6) is immediate. Let us suppose then that

$$P(H_{Q_i}(\tilde{U}_i) \in (\xi, 1-\xi)) < \xi\ .$$

Since $H_{Q_i}(\tilde{U}_i) \in [0,1]$ and $E[H_{Q_i}(\tilde{U}_i)] \in (2\xi, 1-2\xi)$, it follows by Markov's inequality that

$$P(H_{Q_i}(\tilde{U}_i) > 1-\xi) \in \left(\frac{\xi}{1-\xi}, \frac{1-2\xi}{1-\xi}\right) \subseteq (\xi, 1-\xi)\ .$$

Further, there exists a function f such that $\mathbb{1}_{[H_{Q_i}(\tilde{U}_i)>1-\xi]} = f(X_1^N, Y_1^N)$ and $\mathbb{1}_{[H_{R_i}(\tilde{V}_i)>1-\xi]} = f(X_{N+1}^{2N}, Y_{N+1}^{2N})$. It therefore follows from Lemma 7 that

$$P\Big(H_{Q_i}(\tilde{U}_i) > 1-\xi\ ,\ H_{R_i}(\tilde{V}_i) \leq 1-\xi\Big) > \delta(\xi)\ ,$$

implying (6). This completes the proof.

IV. Proof of Theorem 2

Like most proofs of the speed of polarization, our proof of Theorem 2 relies on the following result by Arıkan and Telatar [10], although we need the more general form of the result given in [2, Lemma 2.3].

Lemma 10 ([10],[2])**.** *If Z_n converges almost surely to a random variable Z_∞ and if there exists $K < \infty$ such that*

$$Z_n \leq KZ_{n-1} \quad \text{if } B_n = 0 \tag{7}$$

$$Z_n \leq KZ_{n-1}^2 \quad \text{if } B_n = 1 \tag{8}$$

then

$$\lim_{n\to\infty} P(Z_n < 2^{-2^{n\beta}}) = P(Z_\infty = 0)$$

for all $\beta < 1/2$.

Recall from the proof of Theorem 1 that H_n almost surely converges to a $\{0,1\}$-valued random variable. It then follows from the relations [11]

$$\begin{aligned} Z(A|B)^2 &\leq H(A|B) \\ H(A|B) &\leq \log(1+Z(A|B)) \end{aligned}$$

that Z_n also converges almost surely, and in particular $Z_n \to 0$ whenever $H_n \to 1$, and $Z_n \to 1$ whenever $H_n \to 0$. It then suffices to show that Z_n satisfies inequalities (7) and (8).

We claim that this is indeed the case with $K = 2\psi_0$. To see this, let $\hat{X}_1^{2N}, \hat{Y}_1^{2N}$ be distributed as $P_{X_1^NY_1^N} \cdot P_{X_{N+1}^{2N}Y_{N+1}^{2N}}$, and define the corresponding variables $\hat{U}_i, \hat{V}_i, \hat{Q}_i, \hat{R}_i$ as in (3). We know from [1] that

$$Z(\hat{U}_i+\hat{V}_i|\hat{Q}_i,\hat{R}_i) \leq 2Z(\hat{U}_i|\hat{Q}_i) \tag{9}$$

$$Z(\hat{V}_i|\hat{Q}_i,\hat{R}_i,\hat{U}_i+\hat{V}_i) \leq Z(\hat{U}_i|\hat{Q}_i)^2 \tag{10}$$

Now let (A,B) and $(\hat{A},\hat{B})$ be random variables that can be written as

$$\begin{aligned} (A,B) &= f(X_1^{2N}, Y_1^{2N}) \\ (\hat{A},\hat{B}) &= f(\hat{X}_1^{2N}, \hat{Y}_1^{2N}) \end{aligned}$$

for some function f. Observe that the assumption (1) implies $P_{AB} \leq \psi_0 \cdot P_{\hat{A}\hat{B}}$. Therefore, for binary A we have

$$\begin{aligned} Z(A|B) &= 2\sum_b \sqrt{p_{AB}(0,b)p_{AB}(1,b)} \\ &\leq 2\psi_0 \sum_b \sqrt{p_{\hat{A}\hat{B}}(0,b)p_{\hat{A}\hat{B}}(1,b)} \\ &= \psi_0 \cdot Z(\hat{A}|\hat{B}) \end{aligned} \tag{11}$$

Defining $A = U_i + V_i$ and $B_i = (Q_i, R_i)$ and combining (11) with (9) implies (7) with $K = 2\psi_0$. Similarly, defining $A = V_i$ and $B_i = (Q_i, R_i, U_i + V_i)$ and combining (11) with (10) implies (8) with $K = \psi_0$. This proves Theorem 2 since $\psi_0 < \infty$ by assumption.

	(U_2, U_4)	(U_1, U_3, U_5)	U_6 vs. U_1^5
$S_1 = 0$	$U_4 = 0$		$U_6 \perp U_1^5$
$S_1 = 1$	i.i.d.	$U_5 = U_3$	$U_6 = U_4$
$S_1 = 2$	$U_4 = U_2$		$U_6 \perp U_1^5$
$S_1 = 3$	i.i.d.	$U_5 = U_3 + U_1$	$U_6 = U_4 + U_2$

TABLE I
DISTRIBUTION PROPERTIES OF U_1^6 FOR $N = 8$ AND THE FOUR POSSIBLE INITIAL STATES.

V. PROOF OF THEOREM 3

We give a sketch of the proof, which is divided into two parts. In the first part, we consider $H(U_i|U_1^{i-1}, S_1 = s_1)$. Namely, we assume that the initial state S_1 is know to equal s_1.

Lemma 11. *Consider the stationary Markov process depicted in Figure 1. Then, for $N \geq 8$, the following holds.*

For all $\frac{5N}{8} < i \leq \frac{6N}{8}$ *we have that*

$$H(U_i|U_1^{i-1}, S_1 = s_1) = \begin{cases} 0 & \text{if } s_1 \in \{1, 3\} \\ 1 & \text{if } s_1 \in \{0, 2\} \end{cases} . \quad (12)$$

Proof: The correctness of the lemma is straightforward to validate for $N = 8$ (see the last column of Table I). A simple induction, with $N = 8$ as the basis, is all that is needed for the general case. ■

From the above, we clearly get that for relevant i, $H(U_i|U_1^{i-1}, S_1) = 1/2$, since all 4 states are equally likely as initial states. What remains is to prove that S_1 is essentially known from U_1^{i-1}.

Lemma 12. *Consider the stationary Markov process depicted in Figure 1. Then, there exists an ϵ_N such that*

for all $\frac{5N}{8} < i \leq \frac{6N}{8}$ *we have that*

$$H(S_1|U_1^{i-1}) \leq \epsilon_N \ , \quad \text{and} \quad \lim_{N \to \infty} \epsilon_N = 0 \ . \quad (13)$$

Proof: Consider the first two columns of Table I, and note that U_1^{i-1} encodes $N/8$ i.i.d. realizations of U_1^5. Thus, for N large, the first column allows us to differentiate between $S_1 = 0$, $S_1 = 2$, and $S_1 \in \{1, 3\}$ with very high probability. The second column allows us to differentiate between $S_1 = 1$ and $S_1 = 3$, with very high probability. ■

REFERENCES

[1] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inform. Theory*, vol. 55, pp. 3051–3073, 2009.

[2] E. Şaşoğlu, "Polar coding theorems for discrete systems," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, 2011.

[3] R. Wang, J. Honda, H. Yamamoto, R. Liu, and Y. Hou, "Construction of polar codes for channels with memory," in *Proc. IEEE Inform. Theory Workshop (ITW'2015)*, Jeju Island, Korea, 2015, pp. 187–191.

[4] E. O. Elliot, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J*, vol. 42, pp. 1977–1997, 1968.

[5] E. N. Gilbert, "Capacity of burst-noise channels," *Bell Syst. Tech. J*, vol. 39, pp. 1253–1265, 1960.

[6] M. Mushkin and I. Bar-David, "Capacity and coding for the gilbert-elliot channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1277–1290, 1989.

[7] P. C. Shields, *The Ergodic Theory of Discrete Sample Paths*, ser. Graduate Studies in Mathematics. Providence (R.I.): American Mathematical Society, 1996, vol. 13.

[8] R. G. Gallager, *Information Theory and Reliable Communications*. New York: John Wiley, 1968.

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

[10] E. Arıkan and E. Telatar, "On the rate of channel polarization," in *(ISIT'2009)*, Seoul, South Korea, 2009, pp. 1493–1495.

[11] E. Arıkan, "Source polarization," in *(ISIT'2010)*, Austin, Texas, 2010, pp. 899–903.