Polar Codes for the Deletion Channel: Weak and Strong Polarization

Ido Tal
Technion

Henry D. Pfister
Duke University

Arman Fazeli
UC San Diego

Alexander Vardy
UC San Diego

***Abstract*—This paper presents the first proof of polarization for the deletion channel with a constant deletion rate and a regular hidden-Markov input distribution. A key part of this work involves representing the deletion channel using a trellis and describing the plus and minus polar-decoding operations on this trellis. In particular, the plus and minus operations can be seen as combining adjacent trellis stages to yield a new trellis with half as many stages. Using this viewpoint, we prove a weak polarization theorem for standard polar codes on the deletion channel. To achieve strong polarization, we modify this scheme by adding guard bands of repeated zeros between various parts of the codeword. Using this approach, we obtain a scheme whose rate approaches the mutual information and whose probability of error decays exponentially in the cube-root of the block length.**

I. Introduction

In many communications systems, symbol-timing errors may result in insertion and deletion errors. For example, the *deletion channel* maps a length-N input string to a substring using an i.i.d. process that deletes each input symbol with probabilty δ. These types of channels were first studied in the 1960s [1], [2] and modern coding techniques were first applied to them in [3]. Over the past 15 years, bounds on the capacity of the deletion channel have been significantly improved but a closed-form expression for the capacity remains elusive [4]–[8]. Recently, polar codes were applied to the deletion channel in a series of papers but the question of polarization for non-vanishing deletion rates remained open [9]–[12]. In this work, we show that polar codes can be used to efficiently approach the mutual-information rate between a regular hidden-Markov input process and the output of the deletion channel with constant deletion rate.

In [9], a polar code is designed for the binary erasure channel (BEC) and evaluated on a BEC that also introduces a single deletion. An inner cyclic-redundancy check (CRC) code is used and decoding is performed by running the successive cancellation list (SCL) decoder [13] exhaustively over all compatible erasure locations. The results show one can recover a single deletion in this setting. Extensions to a finite number of deletions are also discussed but the decoding complexity grows faster than N^{d+1}.

In [10], a low-complexity decoder is proposed for the same setup. Its complexity, for a length-N polar code, is roughly $d^2 N \log N$ when d deletions occur. The paper also presents simulation results for polar codes with lengths ranging from 256 to 2048 on two deletion channels. The first channel has a fixed deletion rate of 0.002 and the second introduces exactly 4 deletions. Based on their results, they conjecture that polarization occurs when $N \to \infty$ while the total number of deletions, d, is fixed.

The final papers [11], [12] in this series extend the previous results by proving that weak polarization occurs when $N \to \infty$ and $d = o(N)$. While this result is quite interesting, its proof does not extend to the case of constant deletion rate. For the case where $N \to \infty$ with d fixed, these papers also show strong polarization for the deletion channel and weak polarization for the cascade of deletion channel and a DMC.

In this paper, we combine the well-known trellis representation for channels with synchronization errors [3] with low-complexity joint successive-cancellation decoding for channels with memory [14], [15]. In particular, [3] describes how the input-output mapping of the deletion channel (and other synchronization-error channels) can be represented using a trellis. The main advantage of the trellis perspective is that it naturally generalizes to other channels with synchronization errors (e.g., with insertions, deletions, and errors). The papers [14], [15] describe how the plus and minus polar-decoding operations can be efficiently applied to a channel whose input-output mapping is represented by a trellis. Putting these ideas together defines a low-complexity successive-cancellation decoder for polar codes on the deletion channel that is essentially equivalent to the decoder defined in [10].

Building on previous proofs of polarization for channels with memory [16], [17], this paper also proves weak and strong polarization for the deletion channel. In order to prove strong polarization, guard bands of '0' symbols are embedded in the codewords of Arıkan's standard polar codes. These guard bands allow the decoder to effectively work on independent blocks and enable the proof of strong polarization.

The following theorem is the main result of this paper. We note that the family of allowed input distributions is defined in Subsection II-D, whereas the structure of the codeword is defined in Section VI. Due to space limitations, all proofs are deferred to the extended version [18].

Theorem 1: Fix a regular hidden-Markov input process. For any fixed $\gamma \in (0, 1/3)$, the rate of our coding scheme approaches the mutual-information rate between the input process and the deletion channel output. For large enough blocklength Λ, the decoding error probability is at most $2^{-\Lambda^\gamma}$.

II. Background

A. Notation

The natural numbers are denoted by $\mathbb{N} \triangleq \{1, 2, \ldots\}$. We also define $[m] \triangleq \{1, 2, \ldots, m\}$ for $m \in \mathbb{N}$. Let $\mathcal{X}$ denote

The work of H. D. Pfister was supported in part by the National Science Foundation (NSF) under Grant No. 1718494.

a finite set (e.g., the input alphabet of a channel). In this paper, we fix $\mathcal{X} = \{0, 1\}$ as the binary alphabet. Extensions to non-binary alphabets are straightforward, see for example [19, Chapter 3]. Let $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}^N$ be a vector of length $N = 2^n$. We use $[statement]$ to denote the Iverson bracket which evaluates to 1 if $statement$ is true and 0 otherwise. The concatenation of vectors $\mathbf{y} \in \mathcal{X}^{N_1}$ and $\mathbf{y}' \in \mathcal{X}^{N_2}$ lives in $\mathcal{X}^{N_1+N_2}$ and is denoted by $\mathbf{y} \odot \mathbf{y}'$. The length of a vector $\mathbf{y}$ is denoted by $|\mathbf{y}|$.

In this paper, we use the standard Arıkan transform presented in the seminal paper [20]. Generalization to other kernels [21] is straightforward. The length $N = 2^n$ Arıkan transform of $\mathbf{x} \in \mathcal{X}^N$, is defined recursively using length-$N/2$ binary vectors, $\mathbf{x}^{[0]}$ and $\mathbf{x}^{[1]}$:

$$\mathbf{x}^{[0]} \triangleq (x_1 \oplus x_2, x_3 \oplus x_4, \ldots, x_{N-1} \oplus x_N) \,, \tag{1}$$

$$\mathbf{x}^{[1]} \triangleq (\quad x_2, \quad x_4, \ldots, \quad x_N) \,, \tag{2}$$

where $\oplus$ denotes modulo-2 addition. Then, for any sequence $b_1, b_2, \ldots, b_\lambda \in \{0, 1\}$ with $\lambda \leq n$, we extend this notation to define the vector $\mathbf{x}^{[b_1,b_2,\ldots,b_\lambda]} \in \mathcal{X}^{2^{n-\lambda}}$ recursively via

$$\mathbf{z} = \mathbf{x}^{[b_1,b_1,\ldots,b_{\lambda-1}]} \,, \quad \mathbf{x}^{[b_1,b_2,\ldots,b_\lambda]} = \mathbf{z}^{[b_\lambda]}. \tag{3}$$

Specifically, if $\lambda = n$, then the vector $\mathbf{x}^{[b_1,b_2,\ldots,b_\lambda]}$ is a scalar. This scalar is denoted $u_{i(\mathbf{b})}$, where $\mathbf{b}$ defines the index

$$i(\mathbf{b}) \triangleq 1 + \sum_{j=1}^{n} b_j 2^{n-j} \,. \tag{4}$$

The transformed length-N vector is given by

$$\mathbf{u} = (u_1, \ldots, u_N) = \mathcal{A}_n(\mathbf{x}) \,, \tag{5}$$

where $\mathcal{A}_n \colon \mathcal{X}^{2^n} \to \mathcal{X}^{2^n}$ is called the Arıkan transform of order n. Its inverse is denoted $\mathcal{A}_n^{-1}$ and satisfies $\mathcal{A}_n^{-1} = \mathcal{A}_n$.

B. Deletion Channel

Let $W(\mathbf{y}|\mathbf{x})$ denote the transition probability of N uses of the deletion channel with constant deletion rate δ. The input is denoted by $\mathbf{x} \in \mathcal{X}^N$ and the output $\mathbf{y}$ has a random length $M = |\mathbf{y}|$ supported on $\{0, 1, \ldots, N\}$. This channel is equivalent to a BEC with erasure probability δ followed by a device that removes all erasures from the output. Thus, $W(\mathbf{y}|\mathbf{x})$ equals the probability that $N - M$ deletions have occurred, which is $(1-\delta)^M \cdot \delta^{N-M}$, times the number of distinct deletion patterns that produce $\mathbf{y}$ from $\mathbf{x}$, see [4, Section 2].

We will also consider a trimmed deletion channel (TDC) whose output, denoted $\mathbf{y}^*$, is formed by removing all leading and trailing '0' symbols from the deletion channel output $\mathbf{y}$.

C. Trellis Definition

A *trellis* $\mathcal{T}$ is a labeled weighted directed graph $(\mathcal{V}, \mathcal{E})$ whose vertices $\mathcal{V}$ can be arranged into a sequence of sets such that the edges $\mathcal{E}$ only connect adjacent sets. Each edge $e \in \mathcal{E}$ has a weight $w(e) \in \mathbb{R}$ and a label $\ell(e) \in \mathcal{X}$. A *trellis section* comprises two adjacent sets of vertices along with the edges that connect them. See Fig. 1 for an example with 4 sections. The weight of a path through the trellis is defined as the product of the weights on each edge in the path times the weights of the initial and final vertices (denoted $q(s)$ and $r(s)$, respectively). Thus, an N-section trellis naturally defines a *path-sum* function $T \colon \mathcal{X}^N \to \mathbb{R}$, where $T(\mathbf{x})$ equals the sum of the path weights over all paths whose length-N label sequences match $\mathbf{x}$.

Let $\mathcal{T}$ be a trellis with $N = 2^n$ distinct sections. For the j-th section with $j \in [N]$, let $\mathcal{V}_{j-1} \subseteq \mathcal{V}$ be the set of starting states, $\mathcal{V}_j \subseteq \mathcal{V}$ be the set of ending states, and $\mathcal{E}_j$ be the set of edges that connect these states. For an edge $e \in \mathcal{E}$, we denote the starting and ending states by $\sigma(e)$ and $\tau(e)$, respectively. For trellis $\mathcal{T}$, we write the path-sum function $T \colon \mathcal{X}^N \to \mathbb{R}$ as

$$T(\mathbf{x}) \triangleq \sum_{\substack{e_1 \in \mathcal{E}_1, \\ \ell(e_1) = x_1}} \sum_{\substack{e_2 \in \mathcal{E}_2, \\ \ell(e_2) = x_2}} \cdots \sum_{\substack{e_N \in \mathcal{E}_N, \\ \ell(e_N) = x_N}} q(\sigma(e_1))\, r(\tau(e_N)) \times \prod_{j=1}^{N} w(e_j) \times \prod_{j=1}^{N-1} [\tau(e_j) = \sigma(e_{j+1})] \,.$$

where $q \colon \mathcal{V}_0 \to \mathbb{R}$ is the initial state probability and $r \colon \mathcal{V}_N \to \mathbb{R}$ encapsulates prior knowledge about the final channel state.

D. FAIM processes

For simplicity, this paper sometimes emphasizes the uniform (i.e., i.i.d. Bernoulli $1/2$) input distribution. However, this input distribution is known to be sub-optimal in terms of information rate for the deletion channel [4], [6]–[8]. Thus, one stands to benefit by considering a larger class of input distributions.

Towards this end, let $\mathcal{S}$ be a given finite set. Each element of $\mathcal{S}$ is a state of an input process. In the following[1] definition, we have for all $j \in \mathbb{Z}$ that $S_j \in \mathcal{S}$ and $X_j \in \mathcal{X}$.

Definition 1 (FAIM process): A strictly stationary process (S_j, X_j), $j \in \mathbb{Z}$ is called a *finite-state, aperiodic, irreducible, Markov* (FAIM) process if, for all j,

$$P_{S_j, X_j | S_{-\infty}^{j-1}, X_{-\infty}^{j-1}} = P_{S_j, X_j | S_{j-1}} \,, \tag{6}$$

is independent of j and the sequence (S_j), $j \in \mathbb{Z}$ is a finite-state Markov chain that is stationary, irreducible, and aperiodic.

For a FAIM process, consider the sequence X_j, for $j \in \mathbb{Z}$. In principle, the distribution of this sequence can be computed by marginalizing the states of the FAIM process (S_j, X_j). Such a sequence is typically called a *hidden-Markov process*. In this paper, we sometimes add the term *regular* to emphasize that the hidden state process is a regular finite-state Markov chain.

III. Trellis representation of joint probability

Consider a vector channel with random input $\mathbf{X} \in \mathcal{X}^N$ and random output $\mathbf{Y}$. For some such channels, the transition probability $\Pr(\mathbf{Y} = \mathbf{y} \,|\, \mathbf{X} = \mathbf{x})$ can be computed efficiently on a trellis with N sections. Likewise, for a regular hidden-Markov input distribution, the function $P_{\mathbf{X}}(\mathbf{x})$ can be computed efficiently on a trellis with N sections. In this paper, we assume the input distribution and channel trellises are combined into a single trellis that is used to represent the entire joint probability $\Pr(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}) = P_{\mathbf{X}}(\mathbf{x}) W(\mathbf{y}|\mathbf{x})$.

A. Trellis for deletion channel with i.i.d. input

This trellis representation for the deletion channel can also be found in [3]. Since the deletion channel is not memoryless, it can be beneficial to use an input distribution with memory [5]–[8]. For simplicity, we restrict our description to an i.i.d. Bernoulli input distribution $P_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{N} P_X(x_j)$.

[1] The definition of FAIM and FAIM-derived processes here is a specialization of the definition given in [16].

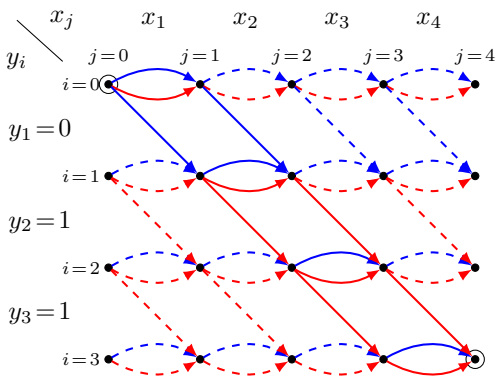

Fig. 1. A trellis for the binary deletion channel corresponding to a codeword length of $N=4$ and received word $\mathbf{y}=(011)$ of length $M=3$. Vertices are denoted $v_{i,j}$ with $0\le i\le M$ and $0\le j\le N$. All blue edges have label '0' while all red edges have label '1'. The horizontal edges are weighted by the probability $\delta/2$. Diagonal edges are weighted by the probability $(1-\delta)/2$. The two circled vertices have $q(v_{0,0})=r(v_{M,N})=1$, while all other vertices in $\mathcal{V}_0$ and $\mathcal{V}_N$ have q and r values equal to 0, respectively. Edges that can be pruned without changing $T(\mathbf{x})$ are dashed.

Let $\mathbf{y}$ be the realization of a received output vector $\mathbf{Y}$. We now describe a trellis that encapsulates the connections between possible transmitted $\mathbf{x}$, the deletion patterns that occur, the fixed received vector $\mathbf{y}$, and the joint probability $P_{\mathbf{X}}(\mathbf{x})\cdot W(\mathbf{y}|\mathbf{x})$. Each path in the trellis corresponds to a specific transmitted $\mathbf{x}$ and a specific deletion pattern that is compatible with the received $\mathbf{y}$ (see Fig. 1).

Definition 2 (Base Trellis): For N, δ, P_X, M, and $\mathbf{y}\in\mathcal{X}^M$:

- Trellis vertices are denoted $v_{i-1,j-1}$ for $i\in[M+1]$ and $j\in[N+1]$. So, $\mathcal{V}=\cup_{j=0}^{N}\mathcal{V}_j$ and $\mathcal{V}_j=\{v_{i-1,j}\,|\,i\in[M+1]\}$.
- Each trellis edge $e\in\mathcal{E}$ has two attributes associated with it: the label $\ell(e)\in\mathcal{X}$ and the weight $w(e)\in[0,1]$.
- Vertices $v_{i-1,j-1}$ with $i\in[M+1]$ and $j\in[N+1]$ each have up to three outgoing edges: two 'horizontal' edges, each corresponding to a deletion, and one 'diagonal' edge, corresponding to a non-deletion.
- For $i\in[M+1]$ and $j\in[N]$, there are two edges e,e' from $v_{i-1,j-1}$ to $v_{i-1,j}$. These are the 'horizontal' edges associated with x_j being deleted by the channel. The first is associated with $x_j=0$ and has $\ell(e)=0$ and $w(e)=\delta\cdot p_X(0)$. The second is associated with $x_j=1$ and has $\ell(e')=1$ and $w(e')=(1-\delta)\cdot p_X(1)$.
- For $i\in[M]$ and $j\in[N]$, there is a single edge e from $v_{i-1,j-1}$ to $v_{i,j}$. This 'diagonal' edge represents x_j being observed as y_i. Thus, $\ell(e)=y_i$ and $w(e)=(1-\delta)\cdot p_X(y_i)$ is the probability $x_j=y_i$ is sent and not deleted.
- A valid path through the trellis is a directed path starting at a vertex in $\mathcal{V}_0$ and ending at a vertex in $\mathcal{V}_N$.
- Each valid path has a corresponding $\mathbf{x}\in\mathcal{X}^N$ that equals the concatenation of the edge labels along the path.
- The initial vertex is always $v_{0,0}$ and this is enforced by choosing $q(s)=[s=v_{0,0}]$. The final vertex is always $v_{M,N}$, so we choose $r(s)=[s=v_{M,N}]$.
- The probability of a valid path equals the product of the weights along the edges of the path times the weight of the initial vertex $q(s_0)$, where $s_0\in\mathcal{V}_0$, times the weight of the final vertex $r(s_N)$, where $s_N\in\mathcal{V}_N$.

The following lemma states the key property of the trellis.

Lemma 2: Let $\mathcal{T}$ be the base trellis for N uses of the deletion channel with i.i.d. inputs defined by P_X. Then, for $\mathbf{x}\in\mathcal{X}^N$,

$$T(\mathbf{x})=P_{\mathbf{X}}(\mathbf{x})\cdot W(\mathbf{y}|\mathbf{x})\,. \tag{7}$$

IV. Polarization operations on a trellis

Polar plus and minus transforms for channels with memory were first presented in [14], [15]. For a vector channel with input $\mathbf{x}\in\mathcal{X}^N$, N even, and output $\mathbf{y}$, let $\mathcal{T}$ be a trellis with N sections whose path-sum function satisfies $T(\mathbf{x})=\Pr(\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x})$. For this channel, the polar *minus transform* defines a new path-sum function that depends on $\mathbf{z}=\mathbf{x}^{[0]}=(x_1\oplus x_2,\ldots,x_{N-1}\oplus x_N)$. This new path-sum function is given by

$$\begin{aligned}T^{[0]}(\mathbf{z})&\triangleq\Pr(\mathbf{Y}=\mathbf{y},\mathbf{X}^{[0]}=\mathbf{z})\\&=\sum_{\mathbf{x}\in\mathcal{X}^N}T(\mathbf{x})\prod_{j=1}^{N/2}[x_{2j-1}\oplus x_{2j}=z_j].\end{aligned}$$

Due to the local nature of this reparameterization, there is a modified trellis $\mathcal{T}^{[0]}$ with $N/2$ sections that represents the new path-sum function. Let $\mathcal{E}_t^{[0]}$ denote the edge set for the t-th section of $\mathcal{T}^{[0]}$. Let $\tilde{e}\in\mathcal{E}_t^{[0]}$ be the edge with $\ell(\tilde{e})=z$, $\sigma(\tilde{e})=a$, and $\tau(\tilde{e})=b$. In trellis $\mathcal{T}^{[0]}$, this edge has weight

$$\begin{aligned}w(\tilde{e})=\sum_{\substack{e_1\in\mathcal{E}_{2t-1}:\\ \sigma(e_1)=a}}\;\sum_{\substack{e_2\in\mathcal{E}_{2t}:\\ \tau(e_2)=b}}&w(e_1)\,w(e_2)\\&\times[\tau(e_1)=\sigma(e_2)]\cdot[\ell(e_1)\oplus\ell(e_2)=z].\end{aligned}$$

Definition 3 (Minus Transform): Let $\mathcal{T}$ be a length-N trellis, where N is even. The trellis $\tilde{\mathcal{T}}=\mathcal{T}^{[0]}$ is defined as follows.

- For $0\le\tilde{j}\le N/2$, define $j=2\tilde{j}$.
- The $\tilde{j}$-th set of vertices in $\tilde{\mathcal{T}}$ satisfies $\tilde{\mathcal{V}}_{\tilde{j}}=\mathcal{V}_j$.
- The weight of an edge $\tilde{v}_{\alpha,\tilde{j}}\xrightarrow{\tilde{e}}\tilde{v}_{\gamma,\tilde{j}+1}$ in $\tilde{\mathcal{T}}$ is the sum of the product of the edge weights along each two-step path $v_{\alpha,j}\xrightarrow{e_1}v_{\beta,j+1}\xrightarrow{e_2}v_{\gamma,j+2}$ in $\mathcal{T}$ with $\ell(\tilde{e})=\ell(e_1)\oplus\ell(e_2)$. Such an edge exists in $\tilde{\mathcal{T}}$ if and only if this sum is positive. This implicitly defines the edge set of $\tilde{\mathcal{T}}$.
- The minus operation does not affect initial and final vertices and this implies that $\tilde{q}(s)=q(s)$ and $\tilde{r}(s)=r(s)$.

This lemma states the key property of the minus transform.

Lemma 3: For a length-N trellis $\mathcal{T}$ and $\mathbf{z}\in\mathcal{X}^{N/2}$, we have

$$T^{[0]}(\mathbf{z})=\sum_{\mathbf{x}\in\mathcal{X}^N:\mathbf{x}^{[0]}=\mathbf{z}}T(\mathbf{x})\,.$$

The polar *plus transform* defines a new path-sum function that depends on $\mathbf{x}^{[1]}=(x_2,x_4,\ldots,x_N)$. This is done by using a previously calculated vector $\mathbf{z}\in\mathcal{X}^{N/2}$ and setting $x_{2j-1}=x_{2j}\oplus z_j$ for $j\in[N/2]$. The implied new path-sum function for $\mathbf{x}'\in\mathcal{X}^{N/2}$ is

$$\begin{aligned}T^{[1]}(\mathbf{x}')&\triangleq\Pr(\mathbf{Y}=\mathbf{y},\mathbf{X}^{[1]}=\mathbf{x}',\mathbf{X}^{[0]}=\mathbf{z})\\&=\sum_{\mathbf{x}\in\mathcal{X}^N}T(\mathbf{x})\prod_{j=1}^{N/2}[x_{2j-1}=x_{2j}\oplus z_j]\cdot[x_{2j}=x'_j].\end{aligned}$$

Due to the local nature of this reparameterization, there is a modified trellis $\mathcal{T}^{[1]}$ with $N/2$ sections that represents this new path-sum function. Let $\mathcal{E}_j^{[1]}$ denote the edge set for the

j-th section of $\mathcal{T}^{[1]}$. Let $\tilde{e} \in \mathcal{E}_j^{[1]}$ be the edge with $\ell(\tilde{e}) = x'$, $\sigma(\tilde{e}) = a$, and $\tau(\tilde{e}) = b$. In trellis $\mathcal{T}^{[1]}$, this edge has weight

$$w(\tilde{e}) = \sum_{\substack{e_1 \in \mathcal{E}_{2t-1}: \\ \sigma(e_1)=a}} \sum_{\substack{e_2 \in \mathcal{E}_{2t}: \\ \tau(e_2)=b}} w(e_1)\, w(e_2) \times [\tau(e_1) = \sigma(e_2)] \cdot [\ell(e_1) = z_j \oplus x'] \cdot [\ell(e_2) = x'].$$

Below, the transformed trellis $\mathcal{T}^{[1]}$ is defined in detail for fixed vector $\mathbf{z}$.

Definition 4 (Plus Transform): For N even, let $\mathcal{T}$ be a length-N trellis and $\mathbf{z} \in \mathcal{X}^{N/2}$. Then, trellis $\tilde{\mathcal{T}} = \mathcal{T}^{[1]}$ satisfies:

- For $0 \le \tilde{j} \le N/2$, denote $j = 2\tilde{j}$.
- The vertex set is defined exactly as in the minus transform. The valid starting and ending vertices are unchanged and this implies that $\tilde{q}(s) = q(s)$ and $\tilde{r}(s) = r(s)$.
- The weight of an edge $\tilde{v}_{\tilde{j},\alpha} \xrightarrow{\tilde{e}} \tilde{v}_{\tilde{j}+1,\gamma}$ in $\tilde{\mathcal{T}}$ with label $x \in \mathcal{X}$ is the sum of the weights of all paths $v_{\alpha,j} \xrightarrow{e_1} v_{\beta,j+1} \xrightarrow{e_2} v_{\gamma,j+2}$ in $\mathcal{T}$ such that $x = \ell(e_2)$ and $z_j = \ell(e_1) \oplus \ell(e_2)$. Such an edge exists in $\tilde{\mathcal{T}}$ if and only if this sum is positive. This implicitly defines the edge set of $\tilde{\mathcal{T}}$.

This lemma states the key property of plus transform.

Lemma 4: Let $\mathcal{T}$ be a length N trellis with N even and let $\mathbf{z} \in \mathcal{X}^{N/2}$ be given. Construct $\mathcal{T}^{[1]}$ with respect to fixed vector $\mathbf{z}$. Then, for any $\mathbf{x}' \in \mathcal{X}^{N/2}$, we have

$$T^{[1]}(\mathbf{x}') = T(\mathbf{x}) \ , \quad \text{where } \mathbf{x}^{[0]} = \mathbf{z} \text{ and } \mathbf{x}^{[1]} = \mathbf{x}' \ .$$

Note that the vector $\mathbf{x} \in \mathcal{X}^N$ is uniquely defined by $\mathbf{x}'$ and $\mathbf{z}$.

As in Arıkan's seminal paper [20], the above transforms lead to a successive cancellation decoding algorithm. In brief, given $\mathbf{y}$ we first construct a base trellis $\mathcal{T}$. Then, there is a recursive decoder that, given $\mathcal{T}^{[b_1,b_2,\ldots,b_\lambda]}$, constructs $\mathcal{T}^{[b_1,b_2,\ldots,b_\lambda,0]}$ and calls itself with that argument. When this returns the decoded $\mathbf{x}^{[b_1,b_2,\ldots,b_\lambda,0]}$, it then builds $\mathcal{T}^{[b_1,b_2,\ldots,b_\lambda,1]}$ with respect to those hard decisions and calls itself to decode $\mathbf{x}^{[b_1,b_2,\ldots,b_\lambda,1]}$. Then, the two decoded vectors are combined to form $\mathbf{x}^{[b_1,b_2,\ldots,b_\lambda]}$ and the function returns. The following lemma makes this precise.

Lemma 5: Let $\mathcal{T}$ be a base trellis with $N = 2^n$ sections corresponding to a received word $\mathbf{y}$. For each $i \in [N]$ in order, let $\hat{u}_1^{i-1}$ be a vector of past decisions and $b_1, b_2, \ldots, b_n \in \{0,1\}$ satisfy $i(\mathbf{b}) = i$. Construct $\mathcal{T}^{[b_1,b_2,\ldots,b_n]}$ iteratively as follows. For $\lambda = 1, 2, \ldots, n$, let us define

$$\mathcal{T}^{[b_1,b_2,\ldots,b_\lambda]} \triangleq \begin{cases} (\mathcal{T}^{[b_1,b_2,\ldots,b_{\lambda-1}]})^{[b_\lambda]} & \text{if } \lambda \ge 2 \ , \\ \mathcal{T}^{[b_1]} & \text{if } \lambda = 1. \end{cases}$$

If $b_\lambda = 1$, then we apply the plus transform with respect to the fixed vector

$$\hat{\mathbf{x}}^{[b_1,b_2,\ldots,b_{\lambda-1},0]} = \mathcal{A}_\lambda^{-1}\left(\hat{u}_\tau^\theta\right) \ ,$$

where $\hat{u}_\tau^\theta \triangleq (\hat{u}_\tau, \hat{u}_{\tau+1}, \ldots, \hat{u}_\theta)$ and

$$\theta = \sum_{j=1}^{\lambda} b_j 2^{n-j} \ , \quad \tau = \theta - 2^{n-\lambda} + 1 \ .$$

Then, for $\mathbf{U} = \mathcal{A}_n(\mathbf{X}) \in \mathcal{X}^N$ we have

$$T^{[b_1,b_2,\ldots,b_n]}(u) = \Pr(U_i = u, U_1^{i-1} = \hat{u}_1^{i-1}, \mathbf{Y} = \mathbf{y}) \ .$$

Actually, this lemma is not unique to the deletion channel and it applies to any base trellis for which (7) holds. The above lemma also gives an efficient method for deciding the value of $\hat{u}_i$ at stage i, since

$$\Pr(U_i = u | U_1^{i-1} = \hat{u}_1^{i-1}, \mathbf{Y} = \mathbf{y}) = \frac{T^{[b_1,b_2,\ldots,b_n]}(u)}{\sum_{u' \in \mathcal{X}} T^{[b_1,b_2,\ldots,b_n]}(u')} \ .$$

V. Weak polarization

A key result of this paper is that polar coding schemes can achieve the information rate

$$\mathcal{I} \triangleq \lim_{N \to \infty} \frac{I(\mathbf{X};\mathbf{Y})}{N} \tag{8}$$

of the deletion channel, where $\mathbf{X}$ and $\mathbf{Y}$ depend implicitly on N. This existence of this limit is well-known [2], [5]. In this section, we describe weak polarization to this rate for both the deletion channel and the trimmed deletion channel. As in [20], the proof relies on showing a certain process is submartingale which must converge to 0 or 1.

As a first step, we will shortly define three entropies. These are defined with respect to an input $\mathbf{X}$ of length $N = 2^n$, which has a hidden-Markov input distribution, and $\mathbf{U} = \mathcal{A}_n(\mathbf{X})$. The corresponding output is denoted $\mathbf{Y}$. Recall that S_0 and S_N are the (hidden) states of the input process, just before $\mathbf{X}$ is transmitted and right after $\mathbf{X}$ is transmitted, respectively. Also, we denote by $\mathbf{Y}^*$ the result of trimming all leading and trailing '0' symbols from $\mathbf{Y}$. Then, for a given n and $1 \le i \le N = 2^n$, we define the following (deterministic) entropies:

$$h_i = H(U_i | U_1^{i-1}, \mathbf{Y}) \ , \tag{9}$$

$$\hat{h}_i = H(U_i | U_1^{i-1}, S_0, S_N, \mathbf{Y}) \ , \tag{10}$$

$$h_i^* = H(U_i | U_1^{i-1}, \mathbf{Y}^*) \ . \tag{11}$$

Clearly, $h_i^* \ge h_i \ge \hat{h}_i$ and we note that, in the case of a uniform input distribution, h_i and $\hat{h}_i$ are equal.

Following [20], we show weak polarization by considering a sequence $B_1, B_2, \ldots$ of i.i.d. $\mathrm{Ber}(1/2)$ random variables. For any $n \in \mathbb{N}$, let $J_n = i(B_1, B_2, \ldots, B_n)$ be the random index defined by (4), with B_t in place of b_t. We will study the three related random processes defined for $n \in \mathbb{N}$ by

$$H_n = h_{J_n} \ , \quad \hat{H}_n = \hat{h}_{J_n} \ , \quad H_n^* = h_{J_n}^* \ .$$

The idea is to show that $\hat{H}_n$ is a submartingale, converging to either 0 or 1. From this we will infer that $\hat{H}_n$ and H_n^* must converge to either 0 or 1 as well even though neither are necessarily a submartingales. The precise statement and proof of the weak polarization theorem is deferred to [18].

VI. Strong polarization

To rigorously claim a coding scheme for the deletion channel, one must also show strong polarization. So far, we have been unable to prove strong polarization for the *standard* polar code construction. Thus, we will modify the standard coding scheme.

The basic idea is to use standard polar encoding for the first n_0 stages, and then to add a guard band in the middle of the codeword during each additional encoding stage. That is, we will generate $\Phi = 2^{n-n_0}$ blocks of length $N_0 = 2^{n_0}$ bits drawn *independently* from the hidden-Markov input distribution.

Between each two consecutive blocks, we will have a string of '0' symbols, which we term a guard band. The real trick is to remove these guard bands in a controlled fashion. For example, if this can be done perfectly, then the effect of the guard bands would be to add commas between blocks of length N_0. The received sequence would then be the statistically independent blocks $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_\Phi$, where $\mathbf{Y}_\phi$ is the output of the channel corresponding to the input segment $\mathbf{X}_t = X_{(\phi-1)\cdot N_0+1}^{\phi \cdot N_0}$. For encoding, the Honda-Yamamoto scheme [22] is be applied to Φ independent blocks of information bits.

Since the received blocks are statistically independent, strong polarization should occur after stage n_0. But, this claim a bit subtle because we carry out one process for the first n_0 stages and then switch to another. Hence, we are in a different setting than that considered in the seminal paper on strong polarization, [23]. However, by [24, Lemma 40], we can indeed establish strong polarization (see also [25]).

Our procedure to remove the above guard bands is not perfect but it can be designed to succeed with high probability. Let the transmitted word be $\mathbf{G}_{\rm I} \odot \mathbf{G}_\triangle \odot \mathbf{G}_{\rm II}$, where $\mathbf{G}_\triangle$ is a string of '0' symbols termed the guard band, and $\mathbf{G}_{\rm I}$ and $\mathbf{G}_{\rm II}$ are of equal length. Denote the corresponding parts of the received word by $\mathbf{Y}_{\rm I}$, $\mathbf{Y}_\triangle$, and $\mathbf{Y}_{\rm II}$. As a preliminary step, we will remove from the received word $\mathbf{Y}$ all leading and trailing '0' symbols. Then, we will assume that the middle index (rounding down) in the resulting word originated from a guard band symbol. We will partition the word into two words according to this middle index, and remove all leading and trailing '0' symbols from these two words. A moment's thought reveals that, if our assumption is correct (i.e., the middle index corresponds to a guard band symbol), then the two resulting words are simply $\mathbf{Y}_{\rm I}^*$ and $\mathbf{Y}_{\rm II}^*$. That is, $\mathbf{Y}_{\rm I}$ and $\mathbf{Y}_{\rm II}$, with leading and trailing '0' symbols removed. That is, in effect, we have transmitted $\mathbf{G}_{\rm I}$ and $\mathbf{G}_{\rm II}$ not over a deletion channel, but over the trimmed deletion channel defined earlier. We will apply this procedure recursively for $n - n_0$ stages. If during all the recursive steps the middle index does indeed belong to the corresponding guard band, we will have produced $\mathbf{Y}_1^*, \mathbf{Y}_2^*, \ldots, \mathbf{Y}_\Phi^*$. We note that a trellis corresponding to the TDC channel can be defined similarly to the trellis we have presented for the deletion channel. For full details, see [18].

Next, we describe exactly how guard bands are added. For $\mathbf{x} = \mathbf{x}_{\rm I} \odot \mathbf{x}_{\rm II} \in \mathcal{X}^{2^n}$, where

$$\mathbf{x}_{\rm I} = x_1^{2^{n-1}} \in \mathcal{X}^{2^{n-1}}, \quad \mathbf{x}_{\rm II} = x_{2^{n-1}+1}^{2^n} \in \mathcal{X}^{2^{n-1}}$$

are the first and second halves of $\mathbf{x}$ respectively, we define

$$g(\mathbf{x}) \triangleq \begin{cases} \mathbf{x} & \text{if } n \le n_0 \\ g(\mathbf{x}_{\rm I}) \odot \overbrace{00\ldots0}^{\ell_n} \odot g(\mathbf{x}_{\rm II}) & \text{if } n > n_0, \end{cases} \tag{12}$$

$$\ell_n \triangleq 2^{\lfloor (1-\epsilon)(n-1) \rfloor}, \tag{13}$$

where $\epsilon \in (0, 1/2)$ is a 'small' constant specified later. Then, the channel input with added guard bands is given by $g(\mathbf{x})$.

The following lemma shows that the rate-penalty for transmitting $g(\mathbf{x})$ in place of $\mathbf{x}$ is negligible as n_0 increases.

Lemma 6: Let $\mathbf{x}$ be a vector of length $|\mathbf{x}| = 2^n$. Then,

$$|\mathbf{x}| \le |g(\mathbf{x})| < \left(1 + \frac{2^{-(\epsilon \cdot n_0 + 1)}}{1 - 2^{-\epsilon}}\right) \cdot |\mathbf{x}| \,. \tag{14}$$

Proof Outline for Theorem 1: The full proof is deferred to the extended paper [18]. But, we note here a few details. First, weak polarization for the TDC implies that the TDC has the same proportion of high-entropy and low-entropy indices as the original deletion channel. Next, we show that recursive partitioning can be used to the remove guard bands with high probability. As noted, the guard bands also incur a negligible rate penalty. Finally, block independence allows us to prove strong polarization using known techniques [24], [25].

REFERENCES

[1] R. Gallager, "Sequential decoding for binary channels with noise and synchronization errors," 1961, Lincoln Lab Group Report.

[2] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Problemy Peredachi Informatsii*, vol. 3, no. 4, pp. 18–36, 1967.

[3] M. C. Davey and D. J. MacKay, "Reliable communication over channels with insertions, deletions, and substitutions," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 687–698, 2001.

[4] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probability Surveys*, vol. 6, pp. 1–33, 2009.

[5] Y. Kanoria and A. Montanari, "Optimal coding for the binary deletion channel with small deletion probability," *IEEE Trans. Inform. Theory*, vol. 59, no. 10, pp. 6192–6219, 2013.

[6] M. Rahmati and T. M. Duman, "Upper bounds on the capacity of deletion channels using channel fragmentation," *IEEE Trans. Inform. Theory*, vol. 61, no. 1, pp. 146–156, 2015.

[7] J. Castiglione and A. Kavcic, "Trellis based lower bounds on capacities of channels with synchronization errors," in *Information Theory Workshop*. Jeju, South Korea: IEEE, 2015, pp. 24–28.

[8] M. Cheraghchi, "Capacity upper bounds for deletion-type channels," *Journal of the ACM (JACM)*, vol. 66, no. 2, p. 9, 2019.

[9] E. K. Thomas, V. Y. F. Tan, A. Vardy, and M. Motani, "Polar coding for the binary erasure channel with deletions," *IEEE Communications Letters*, vol. 21, no. 4, pp. 710–713, April 2017.

[10] K. Tian, A. Fazeli, A. Vardy, and R. Liu, "Polar codes for channels with deletions," in *55th Annual Allerton Conference on Communication, Control, and Computing*, 2017, pp. 572–579.

[11] K. Tian, A. Fazeli, and A. Vardy, "Polar coding for deletion channels: Theory and implementation," in *IEEE International Symposium on Information Theory*, 2018, pp. 1869–1873.

[12] ——, "Polar coding for deletion channels," 2018, submitted to IEEE Trans. Inform. Theory.

[13] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2213–2226, May 2015.

[14] R. Wang, R. Liu, and Y. Hou, "Joint successive cancellation decoding of polar codes over intersymbol interference channels," 2014, preprint arXiv:1404.3001.

[15] 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.

[16] B. Shuval and I. Tal, "Fast polarization for processes with memory," *IEEE Trans. Inform. Theory*, vol. 65, no. 4, pp. 2004–2020, April 2019.

[17] E. Şaşoğlu and I. Tal, "Polar coding for processes with memory," *IEEE Trans. Inform. Theory*, vol. 65, no. 4, pp. 1994–2003, April 2019.

[18] I. Tal, H. D. Pfister, A. Fazeli, and A. Vardy, "Polar codes for the deletion channel: Weak and strong polarization," 2019, preprint arXiv:1904.13385.

[19] E. Şaşoğlu, "Polar Coding Theorems for Discrete Systems," Ph.D. dissertation, IC, Lausanne, 2011.

[20] 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.

[21] S. B. Korada, E. Şaşoğlu, and R. Urbanke, "Polar codes: Characterization of exponent, bounds, and constructions," *IEEE Trans. Inform. Theory*, vol. 56, no. 12, pp. 6253–6264, December 2010.

[22] 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.

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

[24] B. Shuval and I. Tal, "Universal polarization for processes with memory," 2018, preprint arXiv:1811.05727v1.

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