Polar codes for $q$-ary channels, $q = 2^r$

Woomyoung Park

Abstract—We study polarization for nonbinary channels with input alphabet of size $q = 2^r$, $r = 2, 3, \ldots$. Using Arkan’s polarizing kernel $H_2$, we prove that virtual channels that arise in the process of polarization converge to $q$-ary channels with capacity $1, 2, \ldots, r$ bits, and that the total transmission rate approaches the symmetric capacity of the channel. This leads to an explicit transmission scheme for $q$-ary channels. The error probability of decoding using successive cancellation behaves as $\exp(-N^\alpha)$, where $N$ is the code length and $\alpha$ is any constant less than 0.5.

I. INTRODUCTION

Polarization was first described by Arkan [1] who constructed binary codes that achieve capacity of symmetric memoryless channels (and “symmetric capacity” of general binary-input channels). The main idea of [1] is to combine the bits of the source sequence using repeated application of the “polarization kernel” $H_2 = (1 \ 0 \ 1)$. The resulting linear code of length $N = 2^n$ has the generator matrix which forms a submatrix of $G_N = B H_2^\otimes n$, where $B$ is a permutation matrix. The choice of the rows of $G_N$ is governed by the polarization of virtual channels for individual bits that arise in the process of channel combining and splitting. Namely, the data bits are written in the coordinates that correspond to near-perfect channels while the other bits are fixed to some values known to both the transmitter and the decoder.

A study of polar codes for channels with nonbinary input was undertaken by Şaşoğlu et al. [2] and Mori and Tanaka [3]. For prime $q$, it suffices to take the kernel $H_2$, while for nonprime alphabets, the kernel is time-varying and not explicit. [2] remarks that the transmission scheme that uses the kernel $H_2$ with modulo-$q$ addition for composite $q$ does not necessarily lead to the polarization of the channels to the two extremes. The authors of [2] suggest several alternatives to the kernel $H_2$ that rely on randomized permutations or, in the case of $q = 2^r$, on multilevel schemes that implement polar coding for each of the bits of the symbol independently, combining them in the decoding procedure.

In this paper we study polarization for channels with input alphabet of size $q = 2^r$, $r = 2, 3, \ldots$. Suppose that the channel is given by a stochastic matrix $W(y|x)$ where $x \in X, y \in \mathcal{Y}, X = \{0, 1, \ldots, q - 1\}$, and $\mathcal{Y}$ is a finite alphabet. Assuming that the channel combining is performed using the kernel $H_2$ with addition modulo $q$, we establish results about the polarization of channels for individual symbols. It turns out that virtual channels for the transmitted symbols converge to one of $r+1$ extremal configurations in which $j$ out of $r$ bits are transmitted near-perfectly while the remaining $r-j$ bits carry almost no information. Moreover, the good bits are always aligned to the right of the transmitted $r$-block, and no other situations arise in the limit. Thus, the extremal configurations for information rates that arise as a result of polarization are easily characterized: they form an upper-triangular matrix as described in Sect. II-B.

Another related work is the paper by Abbe and Telatar [4]. In it, the authors observed multilevel polarization under the $(\mathbb{F}_2^r)^+$ addition. The main result of their paper provides a characterization of extremal points of the region of attainable rates when polar codes are used for each of the $r$ users of a multiple-access channel. This result applies to a single-user channel, but does not lead to an easy description of the extremal configurations.

Polarization of the kind established in this paper was also observed in a concurrent independent work by Sahebi and Pradhan [5].

II. POLARIZATION FOR $q$-ARY CHANNELS

We consider combining of the $q$-ary data under the action of the operator $H_2$, where $q = 2^r, r \geq 2$. Let $W : \mathcal{X} \rightarrow \mathcal{Y}, |\mathcal{X}| = q$ be a discrete memoryless channel (DMC). The symmetric capacity of the channel $W$ equals

$$I(W) \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \frac{1}{q} W(y|x) \log \frac{W(y|x)}{\sum_{x' \in \mathcal{X}} q^{-1} W(y|x')}$$

where the base of the logarithm is 2. Define the combined channel $W_2$ and the channels $W^-$ and $W^+$ by

$$W_2(y_1,y_2|u_1,u_2) = W(y_1|u_1 + u_2) W(y_2|u_2),$$

$$W^-(y_1,y_2|u_1) = \sum_{u_2 \in \mathcal{X}} \frac{1}{q} W_2(y_1,y_2|u_1,u_2),$$

$$W^+(y_1,y_2|u_1|u_2) = \frac{1}{q} W_2(y_1,y_2|u_1,u_2),$$

where $u_1, u_2, y_1, y_2$ are $r$-vectors and $+$ is a modulo-$q$ sum. This transformation can be applied recursively to the channels $W^-, W^+$ resulting in four channels of the form $W_{\delta_1 \delta_2}^{b_1 b_2}, b_1, b_2 \in \{+, -\}$. After $n$ steps we obtain $N = 2^n$ channels $W_N^{(j)}$, $j = 1, \ldots, N$. For the case $q = 2$ it is shown in [1] that as $n$ increases, the channels $W_N^{(j)}$ become either almost perfect or almost completely noisy (polarize). In formal terms, for any $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \left\{ \left\{ b \in \{+, -\}^n : I(W_b) \in (\varepsilon, 1-\varepsilon) \right\} \right\} = 0.$$  (3)
In this paper we extend this result to the case $q = 2^r$, $r > 1$.

As shown in [1], after $n$ steps of the transformation (1)-(2) the channels $W_N^{i(i)}: \mathcal{X} \to \mathcal{Y}^N \times \mathcal{X}^{N-1}, 1 \leq i \leq N$ are given by

$$W_N^{i(i)}(y_N^i, u_i^{i-1}| u_i) = \frac{1}{q} \sum_{n_i+1 \in X^{N-1}} W_N^{i}(y_N^i | u_i^N G_N),$$

where $G_N = B H_2^{\otimes n}$ and $B$ is a permutation matrix. Here we use the shorthand notation for sequences of symbols: for instance, $y_N^i \equiv (y_1^i, y_2^i, \ldots, y_N^i)$, etc.

A. Notation

For any pair of input symbols $x, x' \in \mathcal{X}$, the Bhattacharyya distance between them is

$$Z(W_{(x,x')}) = \sum_{y \in \mathcal{Y}} \sqrt{W(y|x)W(y|x')}$$

where $W_{(x,x')}$ is the channel obtained by restricting the input alphabet of $W$ to the subset $\{x, x'\} \subset \mathcal{X}$. For $x = (x_1 x_2 \ldots x_r) \in \mathcal{X}$ let $w_r(x) = \max \{i : x_i \neq 0\}$ and let $d_r(x, y) = w_r(x - y).

Define the quantity $Z_v(W)$ for $v \in \mathcal{X} \setminus \{0\}$:

$$Z_v(W) = \frac{1}{2^r} \sum_{x \in \mathcal{X}} Z(W_{(x,vx)}).$$

Introduce the $i$th average Bhattacharyya distance of the channel $W$ by

$$Z_i(W) = \frac{1}{2^{r-1}} \sum_{x \in \mathcal{X}_i} Z_v(W)$$

where $i = 1, 2, \ldots, r$ and $\mathcal{X}_i$ is the set $\{v \in \mathcal{X} : w_r(v) = i\}$. Then

$$Z(W) \equiv \frac{1}{2^{2r-1}} \sum_{x \neq x'} Z(W_{(x,x')}) = \frac{1}{2^{2r-1}} \sum_{i=1}^{2^r-1} Z_i(W).$$

Recall the setting of [1] for the evolution of the channel parameters. On the set $\Omega = \{\{,\}^n\}$ of semi-infinite binary sequences define a $\sigma$-algebra $\mathcal{F}$ on $\Omega$ generated by the cylinder sets $S(b_1, \ldots, b_n) = \{\omega \in \Omega : \omega_1 = b_1, \ldots, \omega_n = b_n\}$ for all sequences $(b_1, \ldots, b_n) \in \{\{,\}^n\}$ and for all $n \geq 0$. Consider the probability space $(\Omega, \mathcal{F}, P)$, where $P(S(b_1, \ldots, b_n)) = 2^{-n}, n \geq 0$. Define a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$ where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n, n \geq 1$ is generated by the cylinder sets $S(b_1, \ldots, b_n), b_i \in \{,\}.$

Let $B_i, i = 1, 2, \ldots$ be i.i.d. $\{,\}$-valued random variables with $P(B_i = 1) = P(B_i = 0) = 1/2$. The random channel emerging at time $n$ will be denoted by $W^B$, where $B = (B_1, B_2, \ldots, B_n).$ Thus, $P(W^B = W_i^{N(i)}) = 2^{-n}$ for all $i = 1, \ldots, 2^n$. Let $W = W^B; I_n = I(W^B), Z_{i(x,x')} = Z(W_{(x,x')}), Z_{i,v} = Z_v(W^B)$, and $Z_{i,v} = Z_{i,v}(W^B)$. These random variables are adapted to the above filtration (meaning that $I_n$ etc. are measurable w.r.t. $\mathcal{F}_n$ for every $n \geq 1$).

B. Channel polarization

In this section we state a sequence of results that shows that $q$-ary polar codes based on the kernel $H_2$ can be used to transmit reliably over the channel $W$ for all rates $R < I(W)$. The proofs for the most part are given in [6].

Theorem 1: (a) Let $n \to \infty$. The random variable $I_n$ converges a.e. to a random variable $I_\infty$ with $E(I_\infty) = I(W)$.

(b) For all $i = 1, 2, \ldots, r$

$$\lim_{n \to \infty} Z_{i,n} = Z_{i,\infty} \text{ a.e.,}$$

where the variables $Z_{i,\infty}$ take values 0 and 1. With probability one the vector $(Z_{i,\infty}, i = 1, \ldots, r)$ takes one of the following values:

$$(Z_{1,\infty} = 0, Z_{2,\infty} = 0, \ldots, Z_{r-1,\infty} = 0, Z_{r,\infty} = 0)$$

$$(Z_{1,\infty} = 1, Z_{2,\infty} = 0, \ldots, Z_{r-1,\infty} = 0, Z_{r,\infty} = 0)$$

$$(Z_{1,\infty} = 1, Z_{2,\infty} = 1, \ldots, Z_{r-1,\infty} = 0, Z_{r,\infty} = 0)$$

$$(Z_{1,\infty} = 1, Z_{2,\infty} = 1, \ldots, Z_{r-1,\infty} = 1, Z_{r,\infty} = 0)$$

$$\vdots$$

$$(Z_{1,\infty} = 1, Z_{2,\infty} = 1, \ldots, Z_{r-1,\infty} = 1, Z_{r,\infty} = 1).$$

Let us restate part (b) of this theorem for finite $n$.

Proposition 1: Let $\epsilon, \delta > 0$ be fixed. For $k = 0, 1, \ldots, r$ define disjoint events

$$B_{k,n}(\epsilon) = \left\{ \omega : (Z_{1,n}, Z_{2,n}, \ldots, Z_{r,n}) \in R_k \right\}$$

where $R_k = R_k(\epsilon) \equiv \left( \prod_{l=1}^{k} D_l \right) \times \left( \prod_{l=k+1}^{r} D_0 \right)$ and $D_0 = [0, \epsilon], D_1 = (1 - \epsilon, 1]$. Then $P(P(V_{k=0} B_{k,n}(\epsilon)) \geq 1 - \delta$ starting from some $n = n(\epsilon, \delta)$.

We need the following lemma.

Lemma 1: For a DMC with q-ary input, $I(W)$ and $Z(W)$ are related by

$$I(W) \geq \log \frac{2^r}{1 + \sum_{i=1}^{r} 2^{-i} Z_i(W)}$$

$$I(W) \leq \sum_{i=1}^{r} \sqrt{1 - Z_i(W)^2}.$$
\[ E(I_{\infty}) = I(W). \] Then use (a) and (b) to claim that \( E(\{i : Z_{i,\infty} = 0\}) = \sum_{k=0}^{r} kP(I_{\infty} = k) = I(W). \]

We can say a bit more about the nature of convergence established in this proposition. Let us fix \( k \in \{0, 1, \ldots, r\} \) and define the channel for the \( r - k \) rightmost bits of the transmitted symbol as follows:

\[ W^{[r-k]}(y|x) = \frac{1}{2^k} \sum_{x: x^{k+1}=u} W(y|x), \quad u \in \{0,1\}^{r-k} \]

where \( x = (x_1, x_2, \ldots, x_r). \)

**Lemma 2:** Let \( V : \mathcal{X} \rightarrow \mathcal{Y} \) be a DMC and let \( \delta > 0. \) Suppose that \( (Z_{1,n}(V), Z_{2,n}(V), \ldots, Z_{r,n}(V)) \in \mathcal{R}_k(\varepsilon), \) for some \( 0 \leq k \leq r. \) If \( \varepsilon \) is sufficiently small, then \( I(V^{[r-k]}) \geq r-k-\delta. \)

In particular, it suffices to take \( \varepsilon \leq (2^{k+\delta} - 1)/2^{3r+k-1}. \)

It turns out that the channels for individual bits converge to either perfect or fully noisy channels. If the channel for bit \( j \) is perfect then the channels for all bits \( i, r \geq i > j \) are perfect. If the channel for bit \( i \) is noisy then the channels for all bits \( j, 1 \leq j < i \) are noisy. The total number of near-perfect bits approaches \( I(W). \) This is made formal in the next proposition.

**Proposition 3:** Let \( \Omega_k = \{w : (Z_{1,\infty}, Z_{2,\infty}, \ldots, Z_{r,\infty}) = \} \)

\[ \lim_{n \to \infty} |I_n - I(W^{[r-k]})| = 0. \]

**Proof:** For every \( \omega \in \Omega_k \) we have that \( I_n(\omega) \to r-k. \)

Combining this with the previous lemma and Proposition 2(b), we conclude that for such \( \omega \) also \( I(W_n^{[r-k]}) \to r-k. \)

The concluding claim of this section describes the channel polarization and establishes that the total number of bits sent over almost noiseless channels approaches \( NF(W). \)

**Theorem 2:** For any DMC \( W : \mathcal{X} \rightarrow \mathcal{Y} \) the channels \( W_{n}^{[i]} \) polarize to one of the \( r + 1 \) extremal configurations. Namely, let \( V_i = W_{N}^{[i]} \) and

\[ \pi_{k,n} = \frac{|\{i \in [N] : I(V_i) - k < \delta \} \wedge |I(V_i^{[k]}) - k| < \delta \}|}{N}, \]

where \( \delta > 0, \) then \( \lim_{n \to \infty} \pi_{k,n} = P(I_{\infty} = k) \) for all \( k \in \{0, 1, \ldots, r\}. \) Consequently

\[ \sum_{k=0}^{r} k \pi_{k} \to I(W). \]

This theorem follows directly from Theorem 1 and Propositions 2 and 3. Some examples of convergence to the extremal configurations described by this theorem are given in the end of this paper.

**C. Transmission with polar codes**

Let us describe a scheme of transmitting over the channel \( W \) with polar codes. Take \( \varepsilon > 0 \) and choose a sufficiently large \( n. \) Assume that the length of the code is \( N = 2^n. \) Proposition 1 implies that set \([N], \) apart from a small subset, is partitioned into \( r + 1 \) subsets \( A_{k,n} \) such that for \( j \in A_{k,n} \) the vector \((Z_1(W_{N}^{(j)}), Z_2(W_{N}^{(j)}), \ldots, Z_{r}(W_{N}^{(j)})) \in \mathcal{R}_k(\varepsilon). \) Each

\[ j \in A_{k,n} \] refers to an \( r \)-bit symbol in which \( r-k \) rightmost bits correspond to small values of \( Z_{i}(W_{N}^{(j)}) \). To transmit data over the channel, we write the data bits in these coordinates and encode them using the linear transformation \( G_{N}. \)

More specifically, let us order the coordinates \( j \in [N] \) by the increase of the quantity \( \sum_{i=1}^{r} 2^{i-1} Z_{i}(W_{N}^{(j)}) \) and use these numbers to locate the subsets \( A_{k,n} \). We transmit data by encoding messages \( u_{N} = (u_1, \ldots, u_N) \) in which if \( j \in A_{k,n}, k = 0, \ldots, r-1 \) then the symbol \( u_j \) is taken from the subset of symbols of \( \mathcal{X} \) with the first \( k \) symbols fixed and known to both the encoder and the decoder ([1] calls them frozen bits). In particular, the subset \( A_{r,n} \) is not used to transmit data. A polar codeword is computed as \( x_{N}^{k} = u_{N}^{k} G_{N} \) and sent over the channel.

Decoding is performed using the “successive cancellation” procedure of [1] with the obvious constraints on the symbol values. Namely, for \( j = 1, \ldots, N \) put

\[ \hat{u}_{j} = \begin{cases} u_{j}, & j \in A_{r,n} \\ \arg \max_{x} W_{N}^{(j)}(y_{N}^{j}, \hat{u}_{i-1}^{j-1}|x), & j \in A_{k,n} \end{cases}, \]

where if \( j \in A_{k,n}, k = 0, \ldots, r-1 \) then the maximum is computed over the symbols \( x \in \mathcal{X} \) with the fixed (known) values of the first \( k \) bits.

The error probability of this decoding is estimated in Sect. II-E.

**D. Proof of Theorem 1: an outline**

Part (a) of Theorem 1 follows straightforwardly from [1], [2]. Namely, as shown in [1, Prop. 4], \( I(W^{+}) + I(W^{-}) = 2I(W). \) We note that the proof in [1] uses only the fact that \( u_1, u_2 \) are recoverable from \( x_1, x_2 \) which is true in our case. Hence the sequence \( I_n, n \geq 1 \) forms a bounded martingale. By Doob’s theorem it converges a.e. in \( L^{1}(\Omega, \mathcal{F}, P) \) to a random variable \( I_{\infty} \) with \( E(I_{\infty}) = I(W). \)

To prove part (b) we show that each of the \( Z_{v,n} \)’s converge a.s. to a \((0, 1) \) Bernoulli random variable \( Z_{v,\infty}. \) This convergence occurs in a concerted way in that the limit r.v.’s obey \( Z_{j,\infty} \geq Z_{i,\infty} \) a.e. if \( j < i. \) This is shown by observing that for any fixed \( i = 1, \ldots, r \) and for all \( v \in \mathcal{X} \), the \( Z_{v,n}(W) \) converge to identical copies of a Bernoulli random variable.

1) Convergence of \( Z_{v,n}, v \in \mathcal{X} \): In this section we outline the convergence proof of the Bhattacharyya parameters \( Z_{v,n} \) to Bernoulli random variables. The following relations were proved in [2]: for all \( v \in \mathcal{X} \setminus \{0\} \)

\[ Z_{v}(W^{+}) = Z_{v}(W)^2 \]

\[ Z_{v}(W^{-}) \leq 2Z_{v}(W) + \sum_{\delta \in \mathcal{X} \setminus \{0, v\}} Z_{\delta}(W)Z_{v+\delta}(W). \] (9)

They imply the following lemma.

**Lemma 3:** Let

\[ Z_{\max}^{(j)}(W) = \max_{v \in \mathcal{X}^{j}} Z_{v}(W), \quad j = 1, \ldots, r. \]

Then

\[ Z_{\max}^{(r-j)}(W^{+}) = Z_{\max}^{(r-j)}(W)^2, \quad j = 0, \ldots, r-1. \] (10)
Lemma 4: Let $U_n, n \geq 0$ be a sequence of random variables adapted to a filtration $\mathcal{F}_n$ with the following properties:

(i) $U_n \in [0, 1]$
(ii) $P(U_{n+1} = U_n^2 | \mathcal{F}_n) \geq 1/2$
(iii) $U_n \leq qU_n$ for some $q \in \mathbb{Z}_+$

Then there are events $\Omega_0, \Omega_1$ such that $P(\Omega_0 \cup \Omega_1) = 1$ and $U_n(\omega) \to i$ for $\omega \in \Omega_i, i = 0, 1$.

Proof: (a) First let us rescale the process $U_n$ so that in the neighborhood of zero it has a drift to zero. Let $\beta \in (0, 1)$ be such that

\[ q^{\beta} < 1 < 4/3. \]

Let $X_n = U_n^{\beta}$. Take $\tau(\omega)$ to be the first time when $X_n(\omega) \geq 1/2$. Let $Y_n = X_{\min(n,r)}$. On the event $Y_n \geq 1/2$ we have $Y_n = Y_{n+1}$ or $E(Y_{n+1} - Y_n|\mathcal{F}_n) = 0$

while on the event $Y_n < 1/2$ we have

\[ E(Y_{n+1} - Y_n|\mathcal{F}_n) \leq \frac{1}{2}(Y_{n}^2 - Y_n) + \frac{1}{2}(q^\beta Y_n - Y_n) \]

\[ \leq -\frac{1}{8}Y_n \leq 0. \]

This implies that the sequence $Y_n, n \geq 0$ forms a supermartingale which is bounded between 0 and 1. By the convergence theorem, $Y_n \to Y_\infty$ a.e. and in $L^1(\Omega, \mathcal{F}, P)$, where $Y_\infty$ is a random variable supported on $[0, 1]$. This implies that $EY_\infty \geq EY_n \downarrow EY_\infty$. Further, if $X_0 \in [0, 1/4]$ then (since $EY_0 = EY_0$) $P(Y_\infty \leq 1/2) \leq 2EY_0 \leq 1/2$. (16)

(b) Now we shall prove that $P(Y_\infty \in (\delta, 1/2 - \delta)) = 0$ for any $\delta > 0$. From (ii) it follows that $P(X_{n+1} = X_n^2|\mathcal{F}_n) \geq 1/2$, which implies that $P(Y_{n+1} = Y_n^2|\mathcal{F}_n) \geq 1/2$ on $Y_n < 1/2$ (17) for all $n \geq 0$. Suppose that $Y_\infty$ takes values in $(\delta, 1/2 - \delta)$ with probability $\alpha > 0$. Let $A_n = \{ \omega : Y_n \in (\delta, 1/2 - \delta) \}$. Since $Y_n \to Y_\infty$ a.e., the Egorov theorem implies that there is a subset of probability arbitrarily close to $P(A_n)$ which this convergence is uniform, and thus $P(A_n) \geq \alpha/2$ for all sufficiently large $n$. Therefore $P(|Y_{n+1} - Y_n| \geq \delta^2/2) \geq P(Y_{n+1} = Y_n^2, Y_n \in (\delta, 1/2 - \delta)) \geq \alpha/4$, the last step by (17). This however contradicts the almost sure convergence of $Y_n$.

(c) This implies that $P(Y_\infty < 1/2) = P(Y_n \to 0) = P(U_n \to 0)$. From (16)

\[ P(U_n \to 0) \geq \frac{1}{2} \quad \text{provided that } U_0 \leq \left( \frac{1}{q} \right)^{2\beta}. \] (18)

Moreover, if $U_0 \leq (1/2)^{1/\beta}$ then either $Y_n \to 0$ or $Y_n \geq 1/2$ for some $n$. This translates to

\[ P((U_n \to 0) \text{ or } (U_n \geq (1/2)^{1/\beta} \text{ for some } n)) = 1 \] (19)

provided that $U_0 \leq (1/2)^{1/\beta}$.

(d) Let $\delta > 0$ be such that $q^{\delta} < 1 - \delta$ (depending on $q$ this may require taking a sufficiently small $\beta$). Let $L := [0, (1/2)^{1/\beta}]$ and $R := [1 - \delta, 1]$. Observe that the process $U_n$ cannot move from $L$ to $R$ without visiting $C := ((1/2)^{1/\beta}, 1 - \delta)$. Let $\sigma_1$ be the first time when $U_n \in C$, let $\eta_1$ be the first time after $\sigma_1$ when $U_n \in L \cup R$, let $\sigma_2$ be the first time after $\eta_1$ when $U_n \in C$, etc., $\sigma_1 < \eta_1 < \sigma_2 < \eta_2 < \ldots$. We shall prove that every sample path of the process eventually stays outside $C$, i.e., that for almost all $\omega$ there exists $k = k(\omega) < \infty$ such that $\sigma_k(\omega) = \infty$.

Assume the contrary, i.e., $\lim_{k \to \infty} P(\sigma_k < \infty) = \alpha > 0$ (since $P(\sigma_{k+1} < \infty) < P(\sigma_k < \infty)$, this limit exists.) We have

\[ P(\exists k : \sigma_k = \infty) \geq \sum_{j=1}^{\infty} P(\sigma_j \neq \infty; U_{\sigma_j} \in L; \sigma_{j+1} = \infty) \]

\[ \geq \alpha \sum_{j=1}^{\infty} P(U_{\sigma_j} \in L; \sigma_{j+1} = \infty|\sigma_j \neq \infty). \] (20)

Consider the process $U_n' = U_{\sigma_j+n}$ on the event $\sigma_k < \infty$ (with the measure renormalized by $P(\sigma_k < \infty)$). This process has the same properties (i)-(iii) as $U_n$. Let $J = [\log_2(1/2^{1/\beta})]$, then $x^{2^J} \in L$ for any $x \in C$. Therefore, $P(U_j' \in L) \geq 2^{-J}$ by property (ii). Now consider the process $U_j' + n$ on the event $U_j' \in L$. This process has properties (i)-(iii), so we can use (18) to conclude that for

\[ P(U_{\eta_k} \in L; \sigma_{k+1} = \infty|\sigma_k \neq \infty) \geq 2^{-J+1} \]

uniformly in $k$. But then the sum in (20) is equal to infinity, a contradiction.

(e) The proof is completed by showing that the probability of $U_n$ staying in $R^c = [0, 1]\setminus R$ without converging to zero is zero. We know that almost all trajectories stay outside $C$, so suppose that the process starts in $(0, (1/2)^{1/\beta})$. Then the probability that it enters $L$ in a finite number of steps is uniformly bounded from below (this is shown similarly to (20)), so the probability that it does not go to $L$ is zero.
assume that the process starts in $L$, then by (19) it either goes to zero or enters $C$ with probability one. Together with part (d) this implies that the process that starts in $L$ converges to zero or one with probability one.

Lemma 5: Let $V : \mathcal{X} \to \mathcal{Y}$ be a channel. Let $v, v' \in \mathcal{X}\{\emptyset\}$ be such that $w_t(v) \geq w_t(v')$. For any $\delta > 0$ there exists $\delta > 0$ such that $Z(v) \geq 1 - \delta'$ whenever $Z(v) \geq 1 - \delta$. In particular, we can take $\delta = \delta' - 3$. This lemma is proved by relating $Z(V_{x',x})$ to the Euclidean distance between the vectors $(\sqrt{V(y|x)}, y \in \mathcal{Y})$ and $(\sqrt{V(y|x')}, y \in \mathcal{Y})$ and some geometric arguments [6].

Lemma 6: For all $j = 1, \ldots, r$
\[ Z(j)_{\text{max, } n} \overset{a.e.}{\to} Z(j)_{\text{max, } \infty}, \]
where $Z(j)_{\text{max, } \infty}$ is a Bernoulli random variable supported on $\{0, 1\}$.

Convergence of the rv's $Z_{x,n}$ is established in the following lemma.

Lemma 7: $Z_{x,n} \to Z_{x,\infty}$ a.e., where $Z_{x,\infty}$ is a $(0,1)$-valued random variable whose distribution depends only on the ordered weight $w_t(x)$.

2) Proof of Part (b) of Theorem 1: Lemma 8: For any $i = 1, \ldots, r$, the random variable $Z_i$ converges a.e. to a $(0,1)$-valued random variable $Z_i,\infty$. Moreover, $Z_i,\infty \to i$ implies that $Z_{i-1,\infty} \to 1$ a.e.

The first part follows because all the $Z_v, v \in \mathcal{X}$ converge to identical copies of the same random variable. The second part holds true because Lemma 5 allows us to move from higher-weight symbols $v$ in $Z_v$ to lower-weight symbols $v$.

Now the proof of Theorem 1(b) can be completed as follows. We obtain that $Z_{i,\infty}$ is a $(0,1)$ random variable a.e. and for all $i$, and if $Z_{i,\infty} = 1$ then $Z_{i,\infty} = 1$ for all $1 \leq j < i$. Consider the events $\Psi_{i} = \{\omega : Z_{j,\infty} = i\}$, $i = 0, 1, j = 1, \ldots, r$. We have
\[ \Psi_0^{(1)} \supset \Psi_0^{(2)} \supset \cdots \supset \Psi_0^{(r)}. \]
We need to prove that with probability one, the vector $(Z_{i,\infty}, i = 1, \ldots, r)$ takes one of the values (5). With probability one $Z_{r,\infty} = 1$ or 0. If it is equal to 1 then necessarily $Z_{r-1,\infty} = \cdots = Z_{1,\infty} = 1$. Otherwise $Z_{r,\infty} = 0$. In this case it is possible that $Z_{r-1,\infty} = 1$ (in which case $Z_{r-2,\infty} = \cdots = Z_{1,\infty} = 1$) or $Z_{r-1,\infty} = 0$. Of course
\[ P(\Psi_{0}^{(r-1)} \cup \Psi_{1}^{(r-1)}) = 1, \]
so in particular
\[ P(\Psi_{0}^{(r)} \setminus (\Psi_{0}^{(r-1)} \cup (\Psi_{1}^{(r-1)} \setminus \Psi_{1}^{(r)}))) = 0. \]
If $Z_{r-1,\infty} = 0$ then the possibilities are $Z_{r-2,\infty} = 1$ or 0, up to another event of probability 0, and so on. Thus, the union of the disjoint events given by (5) holds with probability one. These arguments complete the proof.

Remark: Şaşo/g325glu et al. [2] note that the symbols can polarize to states that carry partial information about the transmission. In particular, they give an example of a quaternary-input channel $W : \{0, 1, 2, 3\} \to \{0, 1\}$ with $W(0|0) =$ $W(0|2) = W(1|1) = W(1|3) = 1$. This channel has capacity 1 bit. Computing the channels $W^+$ and $W^-$ we find that they are equivalent to the original channel $W$. The conclusion reached in [2] was that there are nonbinary channels that do not polarize under the action of $H_2$.

We observe that the above channel corresponds to the extremal configuration (10) in (5) (the other two configurations arise with probability 0), and therefore is a stable point of the channel combining operation. It is possible to reach capacity by transmitting the least significant bit of every symbol.

E. Rate of polarization and error probability of decoding

The main result in this part extends the argument of [1] to nonbinary alphabets.

Theorem 3: Let $0 < \alpha < 1/2$. For any DMC $W : \mathcal{X} \to \mathcal{Y}$ with $I(W) > 0$ and any $R < I(W)$ there exists a sequence of $r$-tuples of disjoint subsets $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_{r-N}$ of $[N]$ such that $\sum_k |A_{k,N}|(r-k) \geq NR$ and $Z(W(i)) < 2^{-N}$ for all $i \in A_{k,N}$, all $v \in \bigcup_{k=k+1}^{r}A_i$, and all $k = 0, 1, \ldots, r - 1$. This theorem follows upon combining Lemma 4 and a result of [7].

The next theorem gives an estimate of the block error rate.

Theorem 4: Let $0 < \alpha < 1/2$ and let $0 < R < I(W)$, where $W : \mathcal{X} \to \mathcal{Y}$ is a DMC. The best achievable error probability of block error under successive cancellation decoding at block length $N = 2^n$ and rate $R$ satisfies $P_e = O(2^{-N^n})$.

F. Examples

A simple example is given by the ordered erasure channel, defined as $W_r : \mathcal{F}_r \to \mathcal{F}_r \cup \{?\}^r$, where
\[ W_r(y|x) = \begin{cases} x_0, & y = x, \\ \epsilon_i, & y = (? \ldots ?x_{i+1} \ldots x_r), 1 \leq i \leq r \\ \end{cases} \]
and $W_r(y|x) = 0$ if $y$ does not contain any erased bits and $y \neq x$. Its capacity equals $r - \sum_{i=1}^{r} \epsilon_i$. We computed the behavior of the virtual channels for $n = 15$ steps of the iteration and observed polarization to $r + 1$ levels for several values of $r$.

ACKNOWLEDGMENT: The authors are grateful to Emmanuel Abbe, Eren Şaşo/g325glu, and Emre Telatar (EPFL), and Leonid Korolov, Armand Makowski, and Himanshu Tyagi (UMD) for useful discussions of this work.

REFERENCES

[1] E. Arikit, “Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” IEEE Trans. Inform. Theory, vol. 55, no. 7, pp. 3051–3073, 2009.
[2] E. Şaşo/g325glu, E. Telatar, and E. Arikit, “Polarization for arbitrary discrete memoryless channels,” arXiv:0908.0302.
[3] R. Mori and T. Tanaka, “Channel polarization over q-ary discrete memoryless channels by arbitrary kernels,” in Proc. IEEE Internat. Symp. Inform. Theory, Austin, TX, pp. 894–898, 2010.
[4] E. Abbe and E. Telatar, “Polar codes for the rn-user MAC,” arXiv:1002.0777.
[5] A. G. Sahito and S. S. Pradhan, “Multilevel polarization of polar codes over arbitrary discrete memoryless channels,” Proc. Allerton Conf. 2011.
[6] W. Park and A. Barg, “Polar codes for q-ary channels, q = 2,” arXiv:1107.4965v3, January 2012.
[7] E. Arikit and E. Telatar, “On the rate of channel polarization,” in Proc. IEEE Int. Symp. Inform. Theory, Seoul, Korea, June 28–July 3, 2009, pp. 1493–1495, 2009.