An entropy inequality for \( q \)-ary random variables and its application to channel polarization

Eren Şaşoğlu  
EPFL, Lausanne, Switzerland  
eren.sasoglu@epfl.ch

Abstract—It is shown that given two copies of a \( q \)-ary input channel \( W \), where \( q \) is prime, it is possible to create two channels \( W^- \) and \( W^+ \) whose symmetric capacities satisfy \( I(W^-) \leq I(W) \leq I(W^+) \), where the inequalities are strict except in trivial cases. This leads to a simple proof of channel polarization in the \( q \)-ary case.

Index Terms—Channel polarization, polar codes, entropy inequality.

I. INTRODUCTION AND MAIN RESULT

Arikan’s polar codes [1] are a class of ‘symmetric capacity’-achieving codes for binary-input channels. Their block error probability behaves roughly like \( O(2^{-\sqrt{n}}) \) [2], where \( n \) is the blocklength, and they achieve this performance at an encoding/decoding complexity of order \( N \log N \).

Polar codes for non-binary input channels were considered in [3]. As in the binary case, their construction is based on recursively creating new channels from several copies of the original: Let \( W \) be a discrete memoryless channel with input alphabet \( \mathcal{X} = \{0, \ldots, q-1\} \). Throughout this note, \( q \) will be assumed to be a prime number. The output alphabet \( \mathcal{Y} \) may be arbitrary. We will let \( I(W) \in [0, 1] \) denote the mutual information developed across \( W \) with uniformly distributed inputs [4], i.e.,

\[
I(W) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \frac{1}{q} W(y | x) \log \frac{W(y | x)}{\sum_{x' \in \mathcal{X}} W(y | x')}. 
\]

Let \( X_1, X_2 \) be independent, uniformly distributed inputs to two independent copies of \( W \), and let \( Y_1, Y_2 \) be the corresponding outputs. Consider the one-to-one mapping \( X_1, X_2 \rightarrow U_1, U_2 \)

\[
U_1 = X_1 + X_2 \\
U_2 = X_2, 
\]

where ‘+’ denotes modulo-\( q \) addition. Observe that \( U_1 \) and \( U_2 \) are independent and uniformly distributed over \( \mathcal{X} \). Define the channels

\[
W^- : U_1 \rightarrow Y_1 Y_2, \\
W^+ : U_2 \rightarrow Y_1 Y_2 U_1, 
\]

described through the conditional output probability distributions

\[
W^-(y_1, y_2 | u_1) = \frac{1}{q} \sum_{u_2 \in \mathcal{X}} W(y_1 | u_1 - u_2)W(y_2 | u_2), \\
W^+(y_1, y_2, u_1 | u_2) = \frac{1}{q} W(y_1 | u_1 - u_2)W(y_2 | u_2).
\]

It follows from the chain rule of mutual information that

\[
I(W^-) + I(W^+) = 2I(W). 
\]

Since \( W^- \) and \( W^+ \) are also \( q \)-ary input channels, the above procedure can be applied to each of them, creating the channels \( W^- := (W^-)^-, \; W^- := (W^-)^+, \; W^+ := (W^+)^-, \) and \( W^+ := (W^+)^+ \). Repeating this procedure \( n \) times, one obtains \( 2^n \) channels, \( W^s, \; s \in \{-, +\}^n \), with \( \sum_s I(W^s) = 2^n I(W) \). The main observation that leads the author of [1] to construct polar codes is that these channels are polarized in the following sense:

Theorem 1 ([1],[3]).

\[
\lim_{n \rightarrow \infty} \frac{1}{2^n} \# \{s \in \{-, +\}^n : I(W^s) \in (1 - \delta, 1]\} = I(W), \quad \lim_{n \rightarrow \infty} \frac{1}{2^n} \# \{s \in \{-, +\}^n : I(W^s) \in [0, \delta]\} = 1 - I(W),
\]

for all \( \delta > 0 \).

The proofs given in [1] and [3] for Theorem 1 are based on the following arguments: The symmetric mutual informations of the channels \( W^s \) created by the above procedure have a martingale property, from which it follows that they must converge for almost all paths in the construction. This shows that both limits in Theorem 1 exist. To prove the claim on these limits’ values, it would be sufficient to show that [2] holds with strict inequalities for all \( W^s \), unless \( I(W^s) \in \{0, 1\} \). Observe, however, that since the output alphabets of channels \( W^s \) grow as the construction size increases, this approach would require the aforementioned inequality to hold uniformly for all \( q \)-ary input channels. This difficulty is circumvented in [1] and [3] by appropriately defining an auxiliary channel parameter \( Z(W) \) and proving the convergence of \( Z(W^s) \) to \( \{0, 1\} \) by the above arguments, which then implies the convergence of \( I(W^s) \) to \( \{0, 1\} \).

\footnote{All logarithms in this note will be to the base \( q \).}
The purpose of this note is to provide a proof of Theorem 1 that avoids this indirect approach. In order to do so, we will need the following theorem.

**Theorem 2.** If \( I(W) \in (\delta, 1 - \delta) \) for some \( \delta > 0 \), then there exists an \( \epsilon(\delta) > 0 \) such that

\[
I(W^-) + \epsilon(\delta) \leq I(W) \leq I(W^+) - \epsilon(\delta).
\]

The dependence of \( \epsilon(\delta) \) on the channel \( W \) is only through \( \delta \), and not through particular channel specifications (e.g., output alphabet size).

Theorem 2 will be proved as a corollary to the following lemma, which is the main result reported here.

**Lemma 1.** Let \( X_1, X_2 \in \mathcal{X}, Y_1, Y_2 \in \mathcal{Y} \) be random variables with joint probability density

\[
P_{X_1, Y_1, X_2, Y_2}(x_1, y_1, x_2, y_2) = P_{X_1, Y_1}(x_1, y_1)P_{X_2, Y_2}(x_2, y_2).
\]

If

\[
H(X_1 \mid Y_1), H(X_2 \mid Y_2) \in (\delta, 1 - \delta)
\]

for some \( \delta > 0 \), then there exists an \( \epsilon(\delta) > 0 \) such that

\[
H(X_1 + X_2 \mid Y_1, Y_2) - \max\{H(X_1 \mid Y_1), H(X_2 \mid Y_2)\} \geq \epsilon(\delta).
\]

We will prove Lemma 1 in Section III.

**Proof of Theorem 2.** It suffices to show that \( I(W) - I(W^-) \geq \epsilon(\delta) \), as the equality \( I(W^-) + I(W^+) = 2I(W) \) will then imply the second half of the claim. Let \( X_1, X_2 \in \mathcal{X} \) denote two independent and uniformly distributed inputs to two copies of \( W \), and let \( Y_1, Y_2 \in \mathcal{Y} \) be the corresponding outputs. Since \( W \) is memoryless, \( X_1, X_2, Y_1, Y_2 \) are jointly distributed as in (3). Further, \( I(W) \in (\delta, 1 - \delta) \) implies

\[
1 - I(W) = H(X_1 \mid Y_1) = H(X_2 \mid Y_2) \in (\delta, 1 - \delta).\quad (4)
\]

It then follows from Lemma 1 that

\[
I(W) - I(W^-) = H(X_1 + X_2 \mid Y_1 Y_2) - H(X_1 \mid Y_1) \geq \epsilon(\delta),
\]

completing the proof.

II. PROOF OF THEOREM 1

Let \( B_1, B_2, \ldots \) be \((-+,+)-\) valued i.i.d. random variables with \( \Pr[B_1 = -] = \Pr[B_1 = +] = \frac{1}{2} \). Let \( I_0, I_1, \ldots \) be random variables defined as

\[
I_0 = I(W),
\]

\[
I_n = I(W^{B_1, \ldots, B_n}) \quad n = 1, 2, \ldots
\]

Note that \( I_n \) takes values in \([0, 1]\). Further, it follows from the relation \( I(W^-) + I(W^+) = 2I(W) \) that \( \mathbb{E}[I_{n+1} \mid I_0, \ldots, I_n] = I_n \). Hence, the process \( I_0, I_1, \ldots \) is a bounded martingale, and therefore converges almost surely to a \([0, 1]-\) valued random variable \( I_\infty \). Note, on the other hand, that

\[
\Pr[I_n \in (\delta, 1 - \delta)] = \frac{1}{2^n} \frac{1}{|\{-+, +\}^n|} \mathbb{I}(I(W^n) \in (\delta, 1 - \delta)).
\]

To conclude the proof, it thus suffices to show that \( \Pr[I_\infty = 1] = I(W) \) and \( \Pr[I_\infty = 0] = 1 - I(W) \). To that end, note that the almost sure convergence of \( I_n \) implies \( \mathbb{E}[I_{n+1} - I_n] = \mathbb{E}[I(W^{B_1, \ldots, B_n}) - I(W^{B_1, \ldots, B_n})] \to 0 \). It follows from Theorem 2 that the latter convergence implies \( I_\infty \in \{0, 1\} \) with probability 1. Due to the martingale property of \( I_n \) we have \( \mathbb{E}[I_\infty] = \mathbb{E}[I_0] = I(W) \), from which it follows that \( \Pr[I_\infty = 1] = 1 - \Pr[I_\infty = 0] = I(W) \), completing the proof.

III. PROOF OF LEMMA 1

In what follows, \( H(p) \) and \( H(X) \) will both denote the entropy of a random variable \( X \in \mathcal{X} \) with probability distribution \( p \). We will let \( p_i, i \in \mathcal{X} \) denote the probability distribution with

\[
p_i(m) = p(m - i).
\]

The cyclic convolution of vectors \( p \) and \( r \) will be denoted by \((p \ast r)\). That is,

\[
(p \ast r) = \sum_{i \in \mathcal{X}} p(i)r_i = \sum_{i \in \mathcal{X}} r(i)p_i.
\]

We will also let \( \text{unif}(\mathcal{X}) \) denote the uniform distribution over \( \mathcal{X} \). We will use the following lemmas in the proof:

**Lemma 2.** Let \( p \) be a distribution over \( \mathcal{X} \). Then,

\[
\|p - \text{unif}(\mathcal{X})\|_1 \geq \frac{1}{q \log e} |1 - H(p)|.
\]

**Remark 1.** Lemma 2 partially complements Pinsker’s inequality by providing a lower bound to the \( L_1 \) distance between an arbitrary probability distribution and the uniform distribution by their Kullback–Leibler divergence.

**Proof:**

\[
1 - H(p) = \sum_{i \in \mathcal{X}} p(i) \log \frac{p(i)}{1/q} \\
\leq \log e \sum_{i} p(i) \left[ \frac{p(i) - 1/q}{1/q} \right] \\
\leq q \log e \sum_{i} p(i) |p(i) - 1/q| \\
\leq q \log e \|p - \text{unif}(\mathcal{X})\|_1,
\]

where we used the relation \( \ln t \leq t - 1 \) in the first inequality.

**Remark 2.** Lemma 2 holds for distributions over arbitrary finite sets. That \( |\mathcal{X}| \) is a prime number has no bearing on the above proof.

**Lemma 3.** Let \( p \) be a distribution over \( \mathcal{X} \). Then,

\[
\|p_i \ast p_j\|_1 \geq \frac{1 - H(p)}{2q^2(q - 1) \log e}
\]

for all \( i, j \in \mathcal{X}, i \neq j \). That is, unless \( p \) is the uniform distribution, its cyclic shifts will be separated from each other in the \( L_1 \) distance.
Proof: Let \( j = i + m \) for some \( m \neq 0 \). We will show that there exists a \( k \in \mathcal{X} \) satisfying
\[
|p(k) - p(k + m)| \geq \frac{1 - H(p)}{2q^2(q - 1) \log e},
\]
which will yield the claim since \( \|p_i - p_j\|_1 = \sum_{k \in \mathcal{X}} |p(k) - p(k + m)| \).

Suppose that \( H(p) < 1 \), as the claim is trivial otherwise. Let \( p^{(\ell)} \) denote the \( \ell \)th largest element of \( p \), and let \( S = \{ \ell : p^{(\ell)} \geq \frac{1}{q} \} \). Note that \( S \) is a proper subset of \( \mathcal{X} \). We have
\[
\sum_{\ell = 1}^{[S]} |p^{(\ell)} - p^{(|S|+1)}| = p^{(1)} - p^{(|S|+1)}
\]
\[
\geq \frac{1}{q} - \frac{1 - H(p)}{2q^2(q - 1) \log e}.
\]
Given such an \( \ell \), let \( A = \{1, \ldots, \ell\} \). Since \( q \) is prime, \( \mathcal{X} \) can be written as
\[
\mathcal{X} = \{k, k + m, k + m + m, \ldots, k + m + \ldots + m\}_{q-1 \text{ times}}
\]
for any \( k \in \mathcal{X} \) and \( m \in \mathcal{X} \setminus \{0\} \). Therefore, since \( A \) is a proper subset of \( \mathcal{X} \), there exists a \( k \in A \) such that \( k + m \in A^c \), implying
\[
p(k) - p(k + m) \geq \frac{1 - H(p)}{2q^2(q - 1) \log e},
\]
which yields the claim.

Lemma 4. Let \( p \) and \( r \) be two probability distributions over \( \mathcal{X} \), with \( H(p) \geq \eta \) and \( H(r) \leq 1 - \eta \) for some \( \eta > 0 \). Then, there exists an \( \epsilon_1(\eta) > 0 \) such that
\[
H(p * r) \geq H(r) + \epsilon_1(\eta).
\]

Proof: Let \( e_i \) denote the distribution with a unit mass on \( i \in \mathcal{X} \). Since \( H(p) \geq \eta \Rightarrow H(e_i) = 0 \), it follows from the continuity of entropy that
\[
\min_i \|p - e_i\|_1 \geq \mu(\eta) \tag{5}
\]
for some \( \mu(\eta) > 0 \). On the other hand, since \( H(r) \leq 1 - \eta \), we have by Lemma 3 that
\[
\|p_i - r_j\|_1 \geq \frac{\eta}{2q^2(q - 1) \log e} > 0 \tag{6}
\]
for all pairs \( i \neq j \). Relations (5), (6), and the strict concavity of entropy implies the existence of \( \epsilon_1(\eta) > 0 \) such that
\[
H(p * r) = H \left( \sum_i p(i)r_i \right) \geq \sum_i p(i)H(r_i) + \epsilon_1(\eta) = H(r) + \epsilon_1(\eta).
\]

Proof of Lemma 4: Let \( P_1 \) and \( P_2 \) be two random probability distributions on \( \mathcal{X} \), with
\[
P_1 = P_{X_1|Y_1}(\cdot \mid y_1) \text{ whenever } Y_1 = y_1,
\]
\[
P_2 = P_{X_2|Y_2}(\cdot \mid y_2) \text{ whenever } Y_2 = y_2.
\]
It is then easy to see that
\[
H(X_1 \mid Y_1) = \mathbb{E}[H(P_1)],
\]
\[
H(X_2 \mid Y_2) = \mathbb{E}[H(P_2)],
\]
\[
H(X_1 + X_2 \mid Y_1, Y_2) = \mathbb{E}[H(P_1 * P_2)].
\]
Suppose, without loss of generality, that \( H(X_1 \mid Y_1) \geq H(X_2 \mid Y_2) \). It suffices to show that if \( \mathbb{E}[H(P_1)], \mathbb{E}[H(P_2)] \in (\delta, 1 - \delta) \) for some \( \delta > 0 \), then there exists an \( \epsilon(\delta) > 0 \) such that \( \mathbb{E}[H(P_1 * P_2)] \geq \mathbb{E}[H(P_1)] + \epsilon(\delta) \). To that end, define the event
\[
A = \{H(P_1) > \delta/2, H(P_2) < 1 - \delta/2\}.
\]
Observe that
\[
\delta < \mathbb{E}[H(P_1)] \leq (1 - \mathbb{P}[H(P_1) > \delta/2]) \cdot \delta/2 + \mathbb{P}[H(P_1) > \delta/2],
\]
implying \( \mathbb{P}[H(P_1) > \delta/2] > \frac{\delta}{2 \delta} \). It similarly follows that \( \mathbb{P}[H(P_2) < 1 - \delta/2] < \frac{\delta}{2 \delta} \). Note further that \( H(P_1) \) and \( H(P_2) \) are independent since \( Y_1 \) and \( Y_2 \) are. Thus, \( A \) has probability at least \( \frac{\delta}{2 \delta} = : \epsilon_2(\delta) \). On the other hand, Lemma 4 implies that conditioned on \( A \) we have
\[
H(P_1 * P_2) \geq H(P_1) + \epsilon_1(\delta/2) \tag{7}
\]
for some \( \epsilon_1(\delta/2) > 0 \). Thus,
\[
\mathbb{E}[H(P_1 * P_2)]
\]
\[
= \mathbb{P}[A] \cdot \mathbb{E}[H(P_1 * P_2) \mid A] + \mathbb{P}[A^c] \cdot \mathbb{E}[H(P_1 * P_2) \mid A^c]
\]
\[
\geq \mathbb{P}[A] \cdot \mathbb{E}[H(P_1) + \epsilon_1(\delta/2) \mid A] + \mathbb{P}[A^c] \cdot \mathbb{E}[H(P_1) \mid A^c]
\]
\[
\geq \mathbb{E}[H(P_1)] + \epsilon_1(\delta/2) \epsilon_2(\delta),
\]
where in the first inequality we used (7) and the relation \( H(p * r) \geq H(p) \). Setting \( \epsilon(\delta) := \epsilon_1(\delta/2) \epsilon_2(\delta) \) yields the result. ■
IV. DISCUSSION

The proof of Theorem 2 does not extend trivially to the case of composite input alphabet sizes. In particular, that the cyclic group \((\{0, \ldots, q - 1\}, +)\) is generated by each of its non-zero elements is crucial to the proof of Lemma 3. On the other hand, a weaker statement holds when the input alphabet size is composite: Consider replacing the mapping (1) with

\[
U_1 = X_1 + X_2, \\
U_2 = \pi(X_2),
\]

(8)

where \(\pi\) is a permutation over \(\mathcal{X}\), and define the channels \(W^-: U_1 \rightarrow Y_1 Y_2\) and \(W^+: U_2 \rightarrow Y_2 Y_2 U_1\) accordingly. Then, it can be shown that there exists a permutation \(\pi\) for which Theorem 2 holds, irrespective of the input alphabet size. The proof of this statement is similar to that of Theorem 2 and therefore is omitted. It then follows that channels with composite input alphabet sizes can be polarized in the sense of Theorem 1 if the mapping in (8) is chosen appropriately at each step of construction. Whether such channels can be polarized by recursive application of a fixed mapping is an open question.

ACKNOWLEDGMENT

I would like to thank Emre Telatar for helpful discussions.

REFERENCES

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

[2] E. Arıkan and E. Telatar, “On the rate of channel polarization,” in Proc. 2009 IEEE Int. Symp. Inform. Theory, (Seoul, Korea), pp. 1493–1495, 28 June – 3 July 2009.

[3] E. Şaşoğlu, E. Telatar, and E. Arıkan, “Polarization for arbitrary discrete memoryless channels,” Aug. 2009. [Online]. Available: arXiv:0908.0302 [cs.IT].