An Alternative Proof of Channel Polarization for Channels with Arbitrary Input Alphabets

Jing Guo
University of Cambridge
jg582@cam.ac.uk

Jossy Sayir
University of Cambridge
j.sayir@ieee.org

Minghai Qin
HGST Research
minghai.qin@hgst.com

Albert Guillén i Fàbregas
ICREA & Universitat Pompeu Fabra
University of Cambridge
guillen@ieee.org

***Abstract*—We revisit channel polarization for arbitrary discrete memoryless channels. A closed-form expression is derived to characterize the difference between the mutual information of the original channel and the virtual channels after one step of channel transformation when the input alphabet and the operation used in the channel transformation form a monoid. We then provide an alternative proof to the one given in [4] for the channel polarization theorem for arbitrary DMCs when the input alphabet set forms a group. The results reveal the connections between channel polarization and zero-error capacity.**

I. Introduction

Polar codes were proposed in [1] as a coding technique that provably achieves the capacity of symmetric binary-input discrete memoryless channels (B-DMCs) with low encoding and decoding complexity. The analysis of the capacity-achieving property of polar codes is based on the channel polarization theorem, which is summarized as follows: Given a B-DMC, virtual channels between the bits at the input of a linear encoder and the channel output sequence are created, such that the mutual information in each of these channels converges to either zero or one as the block length tends to infinity; the proportion of virtual channels with mutual information close to one converges to the capacity of the original channel. Polar codes, constructed based on this principle, can achieve the channel capacity under successive cancellation (SC) decoding.

In the celebrated work of Arıkan [1], the channel polarization theorem is proved only for B-DMCs. Later, it was generalized to prime-input discrete memoryless channels (DMCs) in [9], to prime power-input DMCs in [5], [6] and to arbitrary DMCs in [2]–[4], [7], [8], [10]. The proofs in [5]–[8], [10] all follow Arıkan's proof technique, which is based on the properties of the Battacharyya parameter and mutual information of virtual channels. In [9], the channel polarization theorem for prime-input DMCs is proved without considering the Battacharyya parameter. Instead, the proof is based on the entropy inequality of virtual channels, i.e., the mutual information of the virtual channels is strictly different from that of the original channel. As an extension of [9], the channel polarization theorem is proved in [4] for arbitrary DMCs with input alphabet set forming a quasigroup.

This work has been funded in part by the European Research Council under ERC grant agreement 259663, the Spanish Ministry of Economy and Competitiveness under grant TEC2012-38800-C03-03, and the FP7 Network of Excellence NEWCOM# under grant agreement 318306.

In this paper, we revisit the channel polarization problem for arbitrary DMCs. An alternative proof for the channel polarization theorem for arbitrary DMCs is provided. Similarly to [4], our approach does not consider the Battacharyya parameter. There are two main differences between our proof technique and the one proposed in [4]. First, while [4] proves the entropy inequality of virtual channels by lower bounding the mutual information difference between the original and the "worse" virtual channel, we consider the difference between the "better" virtual channel and the original channel, for which a simple expression is given and bounded away from zero when the input alphabet and the operation used in the channel transformation forms a monoid. Though these two ideas might seem similar, this leads to a new approach for proving the strict inequality. Second, our approach makes use of the properties of Markov chains and the zero-error capacity, without involving distances between probability distributions. Moreover, we show that the extremal channels to which the virtual channels converge have a zero-error capacity equal to their capacity. We note that our proof of channel polarization theorem is restricted to group operations for now, while the stronger results in [4] apply to the wider class of quasigroups.

II. Preliminaries

Throughout this paper, we consider the basic channel transformation described in Fig. 1, where $W : \mathcal{X} \to \mathcal{Y}$ is a DMC with input alphabet set $\mathcal{X} = \{0, \ldots, q-1\}$, output alphabet set $\mathcal{Y}$, and $\oplus$ is a binary operation on the set $\mathcal{X}$. Assume that for all $y \in \mathcal{Y}$ there exists $x \in \mathcal{X}$ such that $W(y|x) > 0$, and for all $x \in \mathcal{X}$ there exists $y \in \mathcal{Y}$ such that $W(y|x) > 0$. Assume U_1 and U_2 are independent random variables with uniform distribution taking values from the set $\mathcal{X}$. According to Fig. 1, we have

$$X_1 = U_1 \oplus U_2, \tag{1}$$

$$X_2 = U_2. \tag{2}$$

Let $W^- : \mathcal{X} \to \mathcal{Y}^2$ be the virtual channel between U_1 and Y_1Y_2, and $W^+ : \mathcal{X} \to \mathcal{Y}^2 \times \mathcal{X}$ be the virtual channel between U_2 and $Y_1Y_2U_1$. W^- and W^+ are synthesized after one channel transformation step (see Fig. 1). After n recursive steps of channel transformation, we can synthesize 2^n virtual channels. We follow some notations used in [1] and let W_n be a random variable that chooses equiprobably from all possible

2^n virtual channels after nth step. Let $I_n = I(W_n)$ be the mutual information of W_n. Moreover, we define two random processes $\{I_n; n \geqslant 0\}$ and $\{W_n; n \geqslant 0\}$. It is proved in [1] that $\{I_n; n \geqslant 0\}$ is a bounded martingale for B-DMCs.

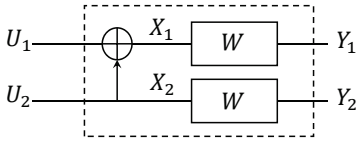

Fig. 1. One step of channel transformation.

The following technical lemma will be used to prove Lemma 4 and Theorem 2.

Lemma 1. *For random variables X, Y, Z whose probability distributions are supported over their respective alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, if X—Y—Z and Y—X—Z form Markov chains, then*

$$\forall x, y, z \text{ such that } P_{XY}(x, y) > 0, P_{Z|Y}(z|y) = P_{Z|X}(z|x). \tag{3}$$

Proof: See Section VI-A. ∎

A consequence of Lemma 1 is that, for any $y \in \mathcal{Y}$, $P_{Z|X}(z|x)$ takes on the same value for all x such that $P_{XY}(xy) > 0$.

We introduce some definitions that will be used throughout this paper.

A. Zero-error capacity

In [11], Shannon introduced the concept of zero-error capacity.

Definition 1 *The zero-error capacity of a noisy channel is the supremum of all rates at which the information can be transmitted with zero error probability.*

Since the capacity of a channel is the supremum of all rates at which the information can be transmitted with vanishing error probability, the zero-error capacity of a channel is always upper bounded by the capacity of this channel. We use $C_0(W)$ to denote the zero-error capacity of a channel W. In this paper, channels with zero zero-error capacity are of primary interest. BECs with strictly positive erasure probability, BSCs with strictly positive crossover probability and AWGN channels with strictly positive noise power are examples of such channels.

Definition 2 *Let $\mathcal{C}_0$ be the set of channels whose zero-error capacity is positive. Let $\mathcal{C}_\emptyset$ be the set of channels whose capacity is zero. Let $\mathcal{C}_0^* = \mathcal{C}_0 \bigcup \mathcal{C}_\emptyset$.*

We claim the following lemma without proof, which is summarized from the statements in [11].

Lemma 2. *For a DMC $W : \mathcal{X} \to \mathcal{Y}$, the following statements are equivalent.*

1) $W \notin \mathcal{C}_0$.
2) $\forall x_1, x_2 \in \mathcal{X}, \sum_{y \in \mathcal{Y}} W(y|x_1)W(y|x_2) > 0$.

B. Basic algebraic structures

We introduce some basic algebraic structures that will be considered in this paper.

Definition 3 *Suppose $\mathcal{X}$ is a set and an operation $\oplus$ is defined over $\mathcal{X}$. $(\mathcal{X}; \oplus)$ forms a monoid if it satisfies the following three axioms.*

1) $\forall x_1, x_2 \in \mathcal{X}, x_1 \oplus x_2 \in \mathcal{X}$.
2) $\forall x_1, x_2, x_3 \in \mathcal{X}, (x_1 \oplus x_2) \oplus x_3 = x_1 \oplus (x_2 \oplus x_3)$.
3) *There exists an element x_0 in $\mathcal{X}$ such that for every element $x \in \mathcal{X}, x \oplus x_0 = x_0 \oplus x = x$. x_0 is also referred as the neutral element of $(\mathcal{X}; \oplus)$.*

In short, a monoid is a single operation algebraic structure satisfying closure, associativity, and the existence of an identity element.

Definition 4 *A group is a monoid in which every element has an inverse.*

For example, the set of numbers $\{0, 1, \cdots, q-1\}$ with multiplication modulo q forms a monoid for all q, but only forms a group for multiplication modulo q if q is prime and 0 is removed from the set.

Definition 5 *Let $(\mathcal{X}; \oplus)$ by any algebraic structure and $\mathcal{X}_s$ be a proper subset of $\mathcal{X}$. If $(\mathcal{X}_s; \oplus)$ forms a group, we call $(\mathcal{X}_s; \oplus)$ a subgroup of $(\mathcal{X}; \oplus)$ and denote this relation by $(\mathcal{X}_s; \oplus) \leqslant (\mathcal{X}; \oplus)$.*

Note that our definition allows for a monoid to have a subgroup.

Definition 6 *Given $(\mathcal{X}_s; \oplus) \leqslant (\mathcal{X}; \oplus)$, for any $x \in \mathcal{X}$, $x \oplus \mathcal{X}_s = \{x \oplus x' \,|\, \forall x' \in \mathcal{X}_s\}$ is called the left coset of $\mathcal{X}_s$ in $\mathcal{X}$ with respect to x, and $\mathcal{X}_s \oplus x = \{x' \oplus x \,|\, \forall x' \in \mathcal{X}_s\}$ is called the right coset of $\mathcal{X}_s$ in $\mathcal{X}$ with respect to x.*

According to Lagrange's Theorem, the left cosets of a group's subgroup partition the group and they have the same cardinality. The left cosets of a monoid's subgroup partition the monoid as well, but their cardinalities can be different.

III. Entropy inequalities for arbitrary DMCs

In this section, we will consider the scenario illustrated in Fig. 1, where U_1, U_2, X_1, X_2 are defined over a finite set $\mathcal{X}$, and the $\oplus$ operation is defined over $\mathcal{X}$ such that $(\mathcal{X}; \oplus)$ forms a monoid.

We first derive a closed-form expression to characterize the difference between the mutual information of the virtual channels and the original channel after a step of channel transformation.

Lemma 3. *Given a DMC $W : \mathcal{X} \to \mathcal{Y}$, we have*

$$I(W^+) - I(W) = I(X_1; Y_1|U_1Y_2). \tag{4}$$

Proof: See Section VI-B. ∎

The rest of the paper is devoted to finding the sufficient and necessary condition for $I(X_1; Y_1|U_1Y_2) > 0$. We first give a sufficient condition.

Lemma 4. *Given a DMC $W : \mathcal{X} \to \mathcal{Y}$, if the channel $W \notin \mathcal{C}_0^*$, then we have*

$$I(X_1; Y_1|U_1Y_2) > 0. \tag{5}$$

Proof: See Section VI-C. ∎

Note that Lemma 4 provides a sufficient but not necessary condition for $I(X_1; Y_1|U_1Y_2) > 0$. Based on Lemma 4, we manage to find a sufficient and necessary condition for $I(X_1; Y_1|U_1Y_2) > 0$, which will be stated in Theorem 2.

Lemma 5. *Given a DMC $W : \mathcal{X} \to \mathcal{Y}$, we have*

$$I(W^-) + I(W^+) \leqslant 2I(W). \tag{6}$$

Proof: See Section VI-D. ∎

The equality in Eq. (6) holds if $(\mathcal{X}; \oplus)$ forms a group. That is, the channel transformation preserves the overall symmetric mutual information when $(\mathcal{X}; \oplus)$ forms a group. Following the same arguments in [1], the random process $\{I_n; n \geqslant 0\}$ is a bounded martingale when the equality in Eq. (6) holds. The sufficient and necessary conditions for the equality in Eq. (6) to hold are studied in [2].

Based on Lemma 3 and Lemma 4 together with Lemma 5, we can prove the main result of the paper, which is the following.

Theorem 1. *For a DMC $W : \mathcal{X} \to \mathcal{Y}$ with $W \notin \mathcal{C}_0^*$,*

$$I(W) - I(W^-) > 0, \tag{7}$$
$$I(W^+) - I(W) > 0. \tag{8}$$

The proof of Theorem 1 is straightforward. First, Eq. (8) is a direct consequence of Lemma 4. Then, Eq. (7) is a direct consequence of Eq. (8) and Lemma 5. Moreover, Lemma 4 will be used in the proof of Lemma 6 (see Section VI-E). Theorem 1 generalizes the results in [9] where Eq. (7) and Eq. (8) are proved for prime-input DMCs only. We will show how Theorem 1 leads to a proof of channel polarization in the next section.

In order to make arguments about entropy inequalities of virtual channels for multiple channel transformation steps, we investigate whether the virtual channels W^- and W^+ inherit the zero zero-error capacity property of the original channel W.

Lemma 6. *Consider a DMC $W : \mathcal{X} \to \mathcal{Y}$. If the channel $W \notin \mathcal{C}_0^*$ and $(\mathcal{X}; \oplus)$ forms a group, then we have*

$$W^+ \notin \mathcal{C}_0^*, \tag{9}$$
$$W^- \notin \mathcal{C}_0^*. \tag{10}$$

Moreover, if $W \notin \mathcal{C}_0$ and $(\mathcal{X}; \oplus)$ only forms a monoid, we have

$$W^+ \notin \mathcal{C}_0, \tag{11}$$
$$W^- \notin \mathcal{C}_0. \tag{12}$$

Proof: See Section VI-E. ∎

IV. Channel polarization for arbitrary DMCs

In the previous section, we proved that the symmetric mutual information of the virtual channels is strictly different from that of the original channel after one step of channel transformation. A natural step forward is to investigate whether the symmetric mutual information of the virtual channels converges asymptotically, and if so, the set of possible values that it converges to.

A. Channel polarization over groups

We first consider the case when $(\mathcal{X}; \oplus)$ forms a group. Since the random process $\{I_n; n \geqslant 0\}$ is a bounded martingale and I_n converges almost everywhere to a random variable I_∞. Then we have

$$\mathsf{E}[|I_{n+1} - I_n|] \to 0. \tag{13}$$

Eq. (13) implies that for any $W \notin \mathcal{C}_0^*$, its corresponding virtual channels will converge to channels in $\mathcal{C}_0^*$ asymptotically.

As for the set of channels the virtual channels will converge to, we need to investigate the set of invariant channels under channel transformation, i.e., channels with $I(X_1; Y_1|U_1Y_2) = 0$.

Definition 7 *Let $\mathcal{C}_{inv}(\mathcal{X})$ denote the the set of channels with input alphabet set $\mathcal{X}$ and $I(X_1; Y_1|U_1Y_2) = 0$ after one step of channel transformation.*

It follows from Lemma 4 that $\mathcal{C}_{inv}(\mathcal{X}) \subset \mathcal{C}_0^*$. The following theorem provides a necessary and sufficient condition for a channel W to be in $\mathcal{C}inv(\mathcal{X})$.

Theorem 2. *Given a DMC $W : \mathcal{X} \to \mathcal{Y}$, a necessary and sufficient condition for $W \in \mathcal{C}_{inv}(\mathcal{X})$ is that both following statements are fulfilled.*

1) *$W : \mathcal{X} \to \mathcal{Y}$ can be decomposed into $t \geqslant 1$ disjoint subchannels $W_i : \mathcal{X}_i \to \mathcal{Y}_i$, with $\mathcal{X}_i \subset \mathcal{X}$ and $\mathcal{Y}_i \subset \mathcal{Y}$ and $W_i \in \mathcal{C}_\emptyset, i \in [t]$, and*
2) *$\exists \mathcal{X}_s \in \{\mathcal{X}_1, \cdots, \mathcal{X}_t\}$ such that $(\mathcal{X}_s; \oplus) \leqslant (\mathcal{X}; \oplus)$ and any $\mathcal{X}_i \in \{\mathcal{X}_1, \cdots, \mathcal{X}_t\}$ is a left coset of $\mathcal{X}_s$.*

Moreover, if $W \in \mathcal{C}_{inv}$, then

$$W^- \in \mathcal{C}_{inv}, \tag{14}$$
$$W^+ \in \mathcal{C}_{inv}. \tag{15}$$

Proof: See Section VI-F. ∎

Theorem 2 implies that the set $\mathcal{C}_{inv}$ is a set of sum channels of which every component channel has zero capacity. Moreover, $\forall W \in \mathcal{C}_{inv}$, the zero-error capacity of channel W equals to its capacity. The logic behind this is as follows: $\forall W \in \mathcal{C}_{inv}$, we have $C_0(W) \leqslant C(W)$, $C(W) = \log(\sum_{i=1}^{t} 2^{C(W_i)}) = \log t$ and $C_0(W) \geqslant \log t = C(W)$. Thus we can conclude that $\forall W \in \mathcal{C}_{inv}$, $C_0(W) = C(W)$. With Eq. (13), Theorem 1, and Theorem 2, we can conclude that successive transformations of channels with zero-error capacity equals to zero will give rise to channels converging towards a set of channels $\mathcal{C}_{inv}(\mathcal{X})$ asymptotically. Moreover, given a channel W whose capacity is larger than its zero-error capacity, successive channel transformations will give rise to channels converging towards a set of channels whose zero-error capacity equals to their capacity.

Now we investigate the limit random variable I_∞. Let W_∞ denote the limit random variable of the random process $\{W_n; n \geqslant 0\}$ as defined previously.

Theorem 3. *Given a DMC $W : \mathcal{X} \to \mathcal{Y}$, W_∞ takes values in the set $\mathcal{C}_{inv}(\mathcal{X})$ and I_∞ takes values in $\left\{\log \frac{|\mathcal{X}|}{|\mathcal{X}_s|} \,|\, \forall(\mathcal{X}_s; \oplus) \leqslant (\mathcal{X}; \oplus)\right\}$.*

For example, given a channel $W : \mathcal{X} \to \mathcal{Y}$ with $|\mathcal{X}| = 6$, I_∞ takes values in $\{\log 1, \log 2, \log 3, \log 6\}$. Theorem 3 is a direct consequence of Theorem 2, so we skip the proof.

B. Channel polarization over monoids

We now briefly discuss channels whose input alphabet set together with $\oplus$ forms a monoid only (not a group). The proof of Theorem 1 is still valid, but the equality in Eq. (6) may not be achieved. A consequence of this is that the random process I_n is no longer a martingale, but a supermartingale instead. Moreover, $I(W^+) - I(W) = I(X_1; Y_1|U_1Y_2) = 0$ does not necessarily imply that $I(W^-) - I(W) = 0$. Instead, $I(W^-) - I(W) = 0$ if $W \in \mathcal{C}_\emptyset$. The possible values of W_∞ and I_∞ are not known. Intuitively, we would guess that W_∞ takes values in $\mathcal{C}_\emptyset$ and $I_\infty = 0$.

V. Conclusion and discussions

In this paper, we have generalized the channel polarization theorem to arbitrary DMCs, using the entropy-based proof proposed in [9]. Furthermore, we have investigated the class of channels that are invariant under channel transformations, which are thus the channels the virtual channels converge to asymptotically. We also revealed some connections between the channel polarization phenomenon and the zero-error capacity. Finally, we discussed channel polarization for channels whose input alphabet set is not a group but only a monoid.

Our overall proof of polarization for arbitrary discrete memoryless channels applies to group operations because we used the existence of a neutral element in the proof of Lemma 4 and the existence of an inverse in the proof of Theorem 2. We know from [4] that an equivalent result holds for the larger class of quasi-groups. Within the framework of our approach, this would suggest that an alternative proof of Lemma 4 not using the neutral element and an alternative proof of Theorem 2 using division instead of the inverse may be possible, but this remains an open problem at this point.

VI. Proofs

In this section, we provide proofs of lemmas and theorems in Section II, Section III and Section IV.

A. Proof of Lemma 1

Since X—Y—Z and Y—X—Z both form Markov chains, we have

$$P_{XZ|Y}(xz|y) = P_{Z|Y}(z|y)P_{X|YZ}(x|yz) \tag{16}$$
$$= P_{Z|Y}(z|y)P_{X|Y}(x|y), \tag{17}$$

and

$$P_{YZ|X}(yz|x) = P_{Y|X}(y|x)P_{Z|XY}(z|xy) \tag{18}$$
$$= P_{Y|X}(y|x)P_{Z|X}(z|x). \tag{19}$$

If $P_{XY}(xy) > 0$, then

$$P_{Z|XY}(z|xy) = P_{Z|X}(z|x) = P_{Z|Y}(z|y). \tag{20}$$

This completes the proof.

B. Proof of Lemma 3

According to the chain rule of entropy, we have

$$I(W^+) - I(W) = I(U_2; Y_1Y_2U_1) - I(U_2; Y_2) \tag{21}$$
$$= H(U_2|Y_2) - H(U_2|Y_1Y_2U_1) \tag{22}$$
$$= I(U_2; Y_1U_1|Y_2) \tag{23}$$
$$= H(U_1Y_1|Y_2) - H(U_1Y_1|U_2Y_2) \tag{24}$$
$$= H(U_1|Y_2) + H(Y_1|U_1Y_2) - H(U_1|U_2Y_2) - H(Y_1|U_1U_2Y_2) \tag{25}$$
$$= H(Y_1|U_1Y_2) - H(Y_1|U_1U_2Y_2) \tag{26}$$
$$= H(Y_1|U_1Y_2) - H(Y_1|X_1) \tag{27}$$
$$= H(Y_1|U_1Y_2) - H(Y_1|X_1U_1Y_2) \tag{28}$$
$$= I(X_1; Y_1|U_1Y_2). \tag{29}$$

In particular, Eq. (26) comes from the fact that $H(U_1|Y_2) = H(U_1|U_2Y_2) = H(U_1)$. Eq. (27) and Eq. (28) come from the fact that Y_1 — X_1— (U_1, Y_2) forms a Markov chain. This completes the proof.

C. Proof of Lemma 4

We prove by contradiction. Assume $I(X_1; Y_1|U_1Y_2) = 0$, we will show that this will lead to a contradiction that $W \in \mathcal{C}_0^*$. By this assumption, we have that X_1 — (U_1, Y_2) — Y_1 forms a Markov chain. By construction (see Fig. 1), (U_1, Y_2) — X_1 — Y_1 forms a Markov chain too. Hence, we are in the scenario of Lemma 1.

If there exists u, y_2 such that $P_{X_1|U_1,Y_2}(x|u, y_2) > 0$ for all $x \in \mathcal{X}$, then by Lemma 1, for any y, $W(y|x)$ has the same value for all $x \in \mathcal{X}$ and hence $I(W) = 0$, which contradicts the condition of the lemma. We can hence assume that, $\forall u, y_2$, the set $\mathcal{X}_{u,y_2} = \{x \in \mathcal{X} \,|\, P_{X_1|U_1,Y_2}(x|u, y_2) > 0\}$ is a proper

subset of $\mathcal{X}$. Consider the set $\mathcal{X}_{0,y_0}$ for some y_0 corresponding to $U_1 = 0$ and $Y_2 = y_0$, where 0 is the neutral element of the monoid $(\mathcal{X}; \oplus)$. Examining Fig. 1 for $U_1 = 0$, we observe that $X_1 = X_2 = U_2$, and hence the setup conditioned on $U_1 = 0$ is equivalent to the setup in Fig. 2. From the figure, it is clear

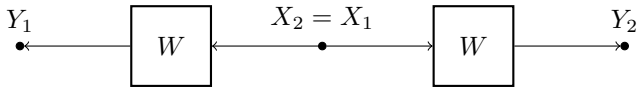

Fig. 2. Setup conditioned on $U_1 = 0$.

that $\mathcal{X}_{0,y_0}$ is non-empty, since otherwise it would contradict the definition of a channel. Since $\mathcal{X}_{0,y_0}$ is a non-empty proper subset of $\mathcal{X}$, its complement $\mathcal{X}^c_{0,y_0} = \mathcal{X} \setminus \mathcal{X}_{0,y_0}$ is also a non-empty proper subset of $\mathcal{X}$. Let x_0 and x_1 be elements of $\mathcal{X}_{0,y_0}$ and $\mathcal{X}^c_{0,y_0}$, respectively. By the definition of $\mathcal{X}_{0,y_0}$, we know that $W(y_0|x_0) > 0$ and $W(y_0|x_1) = 0$. Pick any y_1 such that $W(y_1|x_1) > 0$. Let us assume for now that $W(y_1|x_0) > 0$ as well. $W(y_1|x_0) > 0$ and $W(y_1|x_1) > 0$ imply

$$P_{X_1Y_2}(x_0, y_1) > 0, \text{ and} \tag{30}$$
$$P_{X_1Y_2}(x_1, y_1) > 0, \tag{31}$$

respectively. Lemma 1 with Eq. (30) and Eq. (31) gives, for any y,

$$P_{Y_1|X_1}(y|x_0) = P_{Y_1|X_1}(y|x_1), \tag{32}$$

which is impossible by construction because $W(y_0|x_0) > 0$ and $W(y_0|x_1) = 0$. Hence, our assumption that $W(y_1|x_0) > 0$ leads to a contradiction, and we conclude that $W(y_1|x_0) = 0$.

Having shown that for any $y_1 \in \mathcal{Y}$ such that $W(y_1|x_1) > 0$, $W(y_1|x_0) = 0$, it follows that inputs x_0 and x_1 can be used to transmit 1 bit with zero probability of error over the channel, which contradicts the condition of the lemma. This completes the proof.

D. Proof of Lemma 5

Now we prove that the overall mutual information will be non-increasing after a step of channel transformation. According to the chain rule of mutual information and entropy, we have that

$$\begin{aligned}
I(W^-) + I(W^+) &= I(U_1; Y_1Y_2) + I(U_2; Y_1Y_2U_1) && (33)\\
&= I(U_1; Y_1Y_2) + I(U_2; Y_1Y_2|U_1) && (34)\\
&= I(U_1U_2; Y_1Y_2) && (35)\\
&= I(X_1X_2; Y_1Y_2) && (36)\\
&= H(Y_1Y_2) - H(Y_1Y_2|X_1X_2) && (37)\\
&= H(Y_2) + H(Y_1|Y_2) \\
&\quad - H(Y_1|X_1) - H(Y_2|X_2) && (38)\\
&\leqslant I(X_2; Y_2) + H(Y_1) - H(Y_1|X_1) && (39)\\
&= 2I(W). && (40)
\end{aligned}$$

A sufficient but not necessary condition for the equality in Eq. (39) to hold is that X_1 and X_2 is independent from each other, e.g., $(\mathcal{X}; \oplus)$ forms a group. Studying the full range of operations that yield equality in Eq. (39) is an interesting problem that has been studied in [2].

E. Proof of Lemma 6

We will prove formulas $W^+ \notin \mathcal{C}_0^*$ and $W^- \notin \mathcal{C}_0^*$ when $(\mathcal{X}; \oplus)$ forms a group. It will be clear in the proof that formulas $W^+ \notin \mathcal{C}_0^*$ and $W^- \notin \mathcal{C}_0$ also hold when $(\mathcal{X}; \oplus)$ forms a monoid.

We first prove $W^+ \notin \mathcal{C}_0^*$. The transition probability of channel

$$W^+ : U_2 \to U_1Y_1Y_2$$

is

$$W^+(y_1y_2u_1|u_2) = P_{U_1}(u_1)W(y_1|u_1 \oplus u_2)W(y_2|u_2). \tag{41}$$

Then for any $u_2, u_2' \in \mathcal{X}$, we have

$$\begin{aligned}
&\sum_{y_1\in\mathcal{Y}} \sum_{y_2\in\mathcal{Y}} \sum_{u_1\in\mathcal{X}} W^+(y_1y_2u_1|u_2)W^+(y_1y_2u_1|u_2') && (42)\\
&= \sum_{y_1\in\mathcal{Y}} \sum_{y_2\in\mathcal{Y}} \sum_{u_1\in\mathcal{X}} \big((P_{U_1}(u_1))^2 W(y_2|u_2)W(y_2|u_2') \\
&\quad W(y_1|u_1 \oplus u_2)W(y_1|u_1 \oplus u_2')\big) && (43)\\
&= \left(\underbrace{\sum_{y_2\in\mathcal{Y}} W(y_2|u_2)W(y_2|u_2')}_{>0} \right) \\
&\quad \left(\sum_{u_1\in\mathcal{X}} (P_{U_1}(u_1))^2 \underbrace{\sum_{y_1\in\mathcal{Y}} W(y_1|u_1 \oplus u_2)W(y_1|u_1 \oplus u_2')}_{>0} \right) && (44)\\
&> 0. && (45)
\end{aligned}$$

Eq. (45) along with Lemma 2 implies $C_0(W^+) = 0$, that is to say,

$$W^+ \notin \mathcal{C}_0. \tag{46}$$

Moreover, since

$$\begin{aligned}
I(W^+) &= I(U_2; Y_1Y_2U_1) && (47)\\
&= I(U_2; Y_2) + I(U_2; Y_1U_1|Y_2) && (48)\\
&\geqslant I(W) && (49)\\
&> 0, && (50)
\end{aligned}$$

we have

$$W^+ \notin \mathcal{C}_\emptyset. \tag{51}$$

Based on Eq. (46) and Eq. (51), we conclude that

$$W^+ \notin \mathcal{C}_0^*, \tag{52}$$

which completes the first part of the proof.

The transition probability of the channel

$$W^- : U_1 \to Y_1Y_2$$

is

$$W^-(y_1y_2|u_1) = \sum_{u_2\in\mathcal{X}} P_{U_2}(u_2)W(y_1|u_1 \oplus u_2)W(y_2|u_2). \tag{53}$$

Then for any $u_1, u_1' \in \mathcal{X}$, we have

$$\sum_{y_1\in\mathcal{Y}}\sum_{y_2\in\mathcal{Y}} W^-(y_1y_2|u_1)W^-(y_1y_2|u_1') \tag{54}$$

$$= \sum_{y_1\in\mathcal{Y}}\sum_{y_2\in\mathcal{Y}}\left(\sum_{u_2\in\mathcal{X}} P_{U_2}(u_2)W(y_1|u_1\oplus u_2)W(y_2|u_2)\right.$$
$$\left.\sum_{u_2'\in\mathcal{X}} P_{U_2}(u_2')W(y_1|u_1'\oplus u_2')W(y_2|u_2')\right) \tag{55}$$

$$\geqslant \sum_{y_1\in\mathcal{Y}}\sum_{y_2\in\mathcal{Y}} (P_{U_2}(u_2)W(y_1|u_1\oplus u_2)W(y_2|u_2)$$
$$P_{U_2}(u_2')W(y_1|u_1'\oplus u_2')W(y_2|u_2')) \tag{56}$$

$$= P_{U_2}(u_2)P_{U_2}(u_2')\underbrace{\sum_{y_2\in\mathcal{Y}} W(y_2|u_2)W(y_2|u_2')}_{>0}$$
$$\underbrace{\sum_{y_1\in\mathcal{Y}} W(y_1|u_1\oplus u_2)W(y_1|u_1'\oplus u_2')}_{>0} \tag{57}$$

$$> 0, \tag{58}$$

where Eq. (56) holds for any $u_2, u_2' \in \mathcal{X}$ and this follows from the fact that the summation of non-negative numbers is larger or equal to any addend. Eq. (58) along with Lemma 2 implies that $C_0(W^-) = 0$, that is to say,

$$W^- \notin \mathcal{C}_0. \tag{59}$$

Next we prove $W^- \notin \mathcal{C}_\emptyset$ if $W \notin \mathcal{C}_\emptyset$, i.e., $I(W^-) > 0$ if $I(W) > 0$. We will prove the equivalent proposition that $I(W^-) = 0$ implies $I(W) = 0$. Consider the series of equations, assuming $I(W^-) = 0$, then

$$I(W^+) - I(W) = I(X_1; Y_1|U_1Y_2) \tag{60}$$
$$= I(W) - I(W^-) \tag{61}$$
$$= I(W). \tag{62}$$

This leads to

$$I(X_1; Y_1|U_1Y_2) = I(X_1; Y_1), \tag{63}$$
$$\Rightarrow H(Y_1|U_1Y_2) - H(Y_1|X_1) = H(Y_1) - H(Y_1|X_1), \tag{64}$$
$$\Rightarrow H(Y_1|U_1Y_2) - H(Y_1) = 0, \tag{65}$$
$$\Rightarrow I(Y_1; U_1Y_2) = 0, \tag{66}$$
$$\Rightarrow I(Y_1; U_1) + I(Y_1; Y_2|U_1) = 0, \tag{67}$$
$$\Rightarrow I(Y_1; Y_2|U_1) = 0, \tag{68}$$
$$\Rightarrow I(Y_1; Y_2|U_1 = 0) = 0, \tag{69}$$

where in the last step U_1 is the neutral element of the group. The second equality in Eq. (62) holds when $(\mathcal{X}; \oplus)$ forms a group (see Lemma 5). The left-hand side of Eq. (64) comes from the fact that Y_1—X_1—(U_1, Y_2) forms a Markov chain. Eq. (68) comes from the non-negative property of mutual information. We will look into the joint distribution of (Y_1, Y_2) given $U_1 = 0$. All following arguments are conditioned on $U_1 = 0$ and we omit this expression for simplicity. Since $U_1 = 0$, we let $X = X_1 = X_2 = U_2$ be a uniform random variable on $\mathcal{X}$. Then the joint distribution satisfies

$$P_{Y_1Y_2}(y_1y_2) = \sum_x P_{Y_1Y_2|X}(y_1y_2|x)P_X(x) \tag{70}$$
$$= \sum_x P_{Y_1|X}(y_1|x)P_{Y_2|X}(y_2|x)P_X(x), \tag{71}$$

and the marginal distributions satisfy

$$P_{Y_1}(y_1)P_{Y_2}(y_2) = \left(\sum_x P_{Y_1|X}(y_1|x)P_X(x)\right)$$
$$\left(\sum_x P_{Y_2|X}(y_2|x)P_X(x)\right). \tag{72}$$

Since $I(Y_1; Y_2|U_1 = 0) = 0$, Y_1 and Y_2 are independent (given $U_1 = 0$). We have

$$P_{Y_1Y_2}(y_1y_2) = P_{Y_1}(y_1)P_{Y_2}(y_2), \tag{73}$$
$$\Rightarrow \sum_x P_X(x)P_{Y_1|X}(y_1|x)P_{Y_2|X}(y_2|x)$$
$$= \left(\sum_x P_X(x)P_{Y_1|X}(y_1|x)\right)\left(\sum_x P_X(x)P_{Y_2|X}(y_2|x)\right), \tag{74}$$

$$\Rightarrow \frac{1}{q}\sum_x P_{Y_1|X}(y_1|x)P_{Y_2|X}(y_2|x)$$
$$= \frac{1}{q^2}\left(\sum_x P_{Y_1|X}(y_1|x)\right)\left(\sum_x P_{Y_2|X}(y_2|x)\right), \tag{75}$$

for all $y_1, y_2 \in \mathcal{Y}$. Let $y_1 = y_2 = y$ and note that $P_{Y_1|X}$ and $P_{Y_2|X}$ are both the transition probability of the original channel W, denoted by $P_{Y|X}$, we have

$$\frac{1}{q}\sum_x \left(P_{Y|X}(y|x)\right)^2 = \left(\frac{1}{q}\sum_x P_{Y|X}(y|x)\right)^2. \tag{76}$$

According to Jensen's inequality, the equality is achieved if and only if all terms are equal, i.e., for each $y \in \mathcal{Y}$,

$$P_{Y|X}(y|x) = c \qquad \forall x \in \mathcal{X},$$

where c is a constant depending on y. Thus for each $y \in \mathcal{Y}$, $P_{X|Y}(x|y) = \frac{1}{q}, \forall x \in \mathcal{X}$.

Then $I(W)$ must satisfy

$$I(W) = H(X) - H(X|Y) \tag{77}$$
$$= \log q - \sum_y P_Y(y)H(X|Y = y) \tag{78}$$
$$= \log q - \log q \sum_y P_Y(y) \tag{79}$$
$$= 0. \tag{80}$$

We have shown that $I(W^-) = 0$ implies $I(W) = 0$, equivalently, if $W \notin \mathcal{C}_\emptyset$,

$$W^- \notin \mathcal{C}_\emptyset. \tag{81}$$

Based on Eq. (59) and Eq. (81), we conclude that

$$W^- \notin \mathcal{C}_0^*. \tag{82}$$

Furthermore, we notice that in the proof of Eq. (46) and Eq. (59), the inverse property of a group is not required. Thus if $W \notin \mathcal{C}_0$ and $(\mathcal{X}; \oplus)$ forms a monoid, we have

$$W^- \notin \mathcal{C}_0, \tag{83}$$
$$W^+ \notin \mathcal{C}_0. \tag{84}$$

This completes the proof.

F. Proof of Theorem 2

We first prove the necessary condition for $W \in \mathcal{C}_{inv}$. This is a stronger result than what has been proved in Lemma 4. We follow the idea in the proof of Lemma 4. Let $\mathcal{X}_{u,y_2} = \{x \in \mathcal{X} \,|\, P_{X_1|U_1,Y_2}(x|u,y_2) > 0\}$ and $\mathcal{Y}_{u,y_2} = \{y \in \mathcal{Y} \,|\, P_{Y_1|U_1,Y_2}(y|u,y_2) > 0\}$. Assume $W \in \mathcal{C}_{inv}$, i.e., $I(X_1; Y_1|U_1Y_2) = 0$. According to the proof of Lemma 4, we have following two cases.

Case 1: If $\exists u, y_2$ such that $\mathcal{X}_{u,y_2} = \mathcal{X}$, then $W \in \mathcal{C}_\emptyset$ (see the proof of Lemma 4). Thus Conditions 1) and 2) are fulfilled. Fig. 3 illustrates the channel described by this case, where lines with the same color represent the same transition probability.

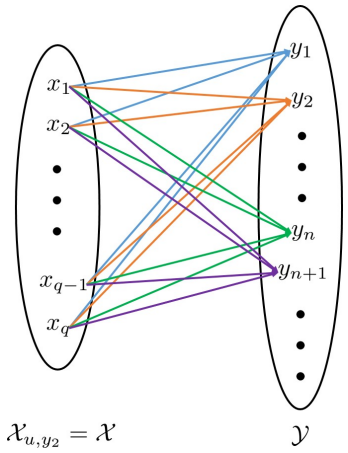

Fig. 3. A channel described by case 1.

Case 2 : If $\forall u, y_2$, the set $\mathcal{X}_{u,y_2}$is a proper subset of $\mathcal{X}$. Examining Fig. 1 for $U_1 = 0$, where 0 is the neutral element for $(\mathcal{X}; \oplus)$, we observe that $X_1 = X_2 = U_2$, and hence the setup conditioned on $U_1 = 0$ is equivalent to the setup in Fig. 2. Fig. 4 illustrates the channel described in this case. We first prove Condition 1). According to the proof of Lemma 4, we have that if $x_0 \in \mathcal{X}_{0,y}$ and $x_1 \in \mathcal{X}_{0,y}^c$, then

$$\sum_{y \in \mathcal{Y}} W(y|x_0)W(y|x_1) = 0. \tag{85}$$

We have

$$\forall y_i, y_j \in \mathcal{Y}, \mathcal{X}_{0,y_i} = \mathcal{X}_{0,y_j} \text{ or } \mathcal{X}_{0,y_i} \cap \mathcal{X}_{0,y_j} = \emptyset. \tag{86}$$

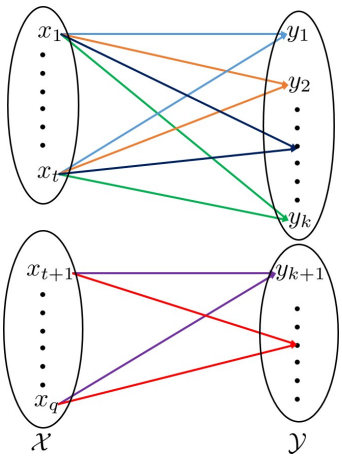

Fig. 4. A channel described by case 2.

This can be seen via a proof by contradiction. Assuming the contrary, we can find $x_0 \in \mathcal{X}_{0,y_i} \cap \mathcal{X}_{0,y_j}$ and $x_1 \in \mathcal{X}_{0,y_i} \cap \mathcal{X}_{0,y_j}^c$ such that

$$W(y|x_0) = W(y|x_1), \tag{87}$$
$$\sum_{y \in \mathcal{Y}} W(y|x_0)W(y|x_1) = 0. \tag{88}$$

Eq. (87) comes from the assumption that $x_0, x_1 \in \mathcal{X}_{0,y_i}$. $x_0 \in \mathcal{X}_{0,y_j}, x_1 \notin \mathcal{X}_{0,y_j}$ together with Eq. (85) leads to Eq. (88). Then we have

$$\sum_{y \in \mathcal{Y}} W(y|x_1)W(y|x_1) = 0, \tag{89}$$

which conflicts with the definition of a channel. Thus, the assumption cannot be true. It follows that $\mathcal{X}$ can be partitioned into t disjoint subsets $\{\mathcal{X}_1, \ldots, \mathcal{X}_t\}$. Let $\mathcal{Y}_i = \{y \in \mathcal{Y} \,|\, \mathcal{X}_{0y} = \mathcal{X}_i\}$. It is easy to show that $\forall x \in \mathcal{X}_i, y \in \mathcal{Y}_i$, $W(y|x) = c_y > 0$ is constant with respect to x. Furthermore, if $x \in \mathcal{X}_i^c, y \in \mathcal{Y}_i$ or $x \in \mathcal{X}_i, y \in \mathcal{Y}_i^c$, then $W(y|x) = 0$. Thus, the channel $W : \mathcal{X} \to \mathcal{Y}$ can be decomposed into t disjoint subchannels $W_i : \mathcal{X}_i \to \mathcal{Y}_i, i \in [t]$ and $W_i \in \mathcal{C}_\emptyset$.

Next we prove Condition 2). Assume that $\mathcal{X}_s \in \{\mathcal{X}_1, \ldots, \mathcal{X}_t\}$ contains the neutral element 0, then $0 \in \mathcal{X}_{0,y_s}$, for all $y_s \in \mathcal{Y}_s$. Notice that for all $x_1, x_2 \in \mathcal{X}_{0,y}, y \in \mathcal{Y}$, $\mathcal{X}_{x_1,y} = \mathcal{X}_{x_2,y}$ since they share the common element $x_1 \oplus x_2$. We have $\forall x_1, x_2 \in \mathcal{X}_{0y_s} = \mathcal{X}_s$, $x_1 \oplus x_2 \in \mathcal{X}_{x_1,y_s} = \mathcal{X}_{x_2,y_s} = \mathcal{X}_s$. Thus, $\mathcal{X}_s$ is closed under $\oplus$. Notice that $\forall x \in \mathcal{X}_s$, $0 \in \mathcal{X}_{x,y_s} = x \oplus \mathcal{X}_s$, thus $\exists x' \in \mathcal{X}_s$ such that $x \oplus x' = 0$. This means that every $x \in \mathcal{X}_s$ has an inverse element $x' \in \mathcal{X}_s$. So we have $(\mathcal{X}_s; \oplus) \leqslant (\mathcal{X}; \oplus)$. Moreover, we note that $\forall x \in \mathcal{X}$, $\exists i \in [t]$ such that $x \oplus \mathcal{X}_s \subset \mathcal{X}_i$. Since the left cosets of a subgroup in a group partitions the group and the subsets $\{\mathcal{X}_1, \ldots, \mathcal{X}_t\}$ partition $\mathcal{X}$, we must have $x \oplus \mathcal{X}_s = \mathcal{X}_i$. Otherwise, there exists $i, j \in [t]$ such that $\mathcal{X}_i \cap \mathcal{X}_j \neq \emptyset$. Thus, we can conclude that the left cosets of $\mathcal{X}_s$ are $\{\mathcal{X}_1, \ldots, \mathcal{X}_t\}$.

Now we prove the sufficiency part. If W fulfills both

conditions, we have

$$
\begin{aligned}
&I(X_1;Y_1|U_1Y_2) && (90)\\
&= H(X_1|U_1Y_2) - H(X_1|U_1Y_1Y_2) && (91)\\
&= \sum_{u_1\in\mathcal{X}}\sum_{y_2\in\mathcal{Y}} P_{U_1Y_2}(u_1y_2)H(X_1|u_1y_2)\\
&\quad - \sum_{u_1\in\mathcal{X}}\sum_{y_1\in\mathcal{Y}}\sum_{y_2\in\mathcal{Y}} P_{U_1Y_1Y_2}(u_1y_1y_2)H(X_1|u_1y_1y_2) && (92)\\
&= \log|\mathcal{X}_s| - \log|\mathcal{X}_s| && (93)\\
&= 0, && (94)
\end{aligned}
$$

where $(\mathcal{X}_s;\oplus) \leqslant (\mathcal{X};\oplus)$. Since any $\mathcal{X}_i \in \{\mathcal{X}_1,\ldots,\mathcal{X}_t\}$ is a left coset of $\mathcal{X}_s$, we have $|\mathcal{X}_i| = |\mathcal{X}_s|$. Eq. (93) follows from the fact that $H(X_1|u_1y_2) = \log|\mathcal{X}_s|$ and $H(X_1|u_1y_1y_2) = \log|\mathcal{X}_s|$.

Furthermore, if $W \in \mathcal{C}_{inv}$, it is easy to verify that both W^- and W^+ can also be decomposed into t disjoint subchannels whose capacity is 0 and the input alphabet set $\mathcal{X}$ has the same partition as W, i.e.,

$$W^- \in \mathcal{C}_{inv}, \tag{95}$$

$$W^+ \in \mathcal{C}_{inv}. \tag{96}$$

This completes the proof.

REFERENCES

[1] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, July 2009.

[2] R. Nasser, "Ergodic theory meets polarization. I: An ergodic theory for binary operations," *arXiv:1406.2943v4 [math.CO]*, 2015.

[3] ——, "Ergodic theory meets polarization. II: A foundation of polarization theory," *arXiv:1406.2949v4 [cs.IT]*, 2015.

[4] R. Nasser and E. Telatar, "Polarization theorems for arbitrary DMCs," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, July 2013, pp. 1297–1301.

[5] W. Park and A. Barg, "Multilevel polarization for nonbinary codes and parallel channels," in *Proc. 49-th Allerton Conf. Commun., Control and Computing*, Monticello, IL, Sept. 2011, pp. 228–234.

[6] ——, "Polar codes for q-ary channels, $q = 2^r$," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Cambridge, MA, July 2012, pp. 2142–2146.

[7] A. G. Sahebi and S. S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels," in *Proc. 49-th Allerton Conf. Commun., Control and Computing*, Monticello, IL, Sept. 2011, pp. 1718 – 1725.

[8] ——, "Multilevel channel polarization for arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7839–7857, Dec. 2013.

[9] E. Şaşoğlu, "An entropy inequality for q-ary random variables and its application to channel polarization," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, June 2010, pp. 1360–1363.

[10] E. Şaşoğlu, E. Telatar, and E. Arıkan, "Polarization for arbitrary discrete memoryless channels," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Oct. 2009, pp. 144–148.

[11] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. 2, no. 3, pp. 8–19, Sept. 1956.