Polar Coding without Alphabet Extension for Asymmetric Models

Junya Honda, *Non-Member,* and Hirosuke Yamamoto, *Fellow, IEEE,*

***Abstract*—This paper considers polar coding for asymmetric settings, that is, channel coding for asymmetric channels and lossy source coding for nonuniform sources and/or asymmetric distortion measures. The difficulty for asymmetric settings comes from the fact that the optimal symbol distributions of codewords are not always uniform. It is known that such nonuniform distributions can be realized by Gallager's scheme which maps multiple auxiliary symbols distributed uniformly to an actual symbol. However, the complexity of Gallager's scheme increases considerably for the case that the optimal distribution cannot be approximated by simple rational numbers. To overcome this problem for the asymmetric settings, a new polar coding scheme is proposed, which can attain the channel capacity without any alphabet extension by invoking results on polar coding for lossless compression. It is also shown that the proposed scheme achieves a better tradeoff between complexity and decoding error probability in many cases.**

***Index Terms*—Asymmetric channels, channel coding, lossy source coding, polar codes.**

I. Introduction

Recently polar coding has attracted much attention for its achievability of Shannon bound with polynomial complexity. Polar codes are originally proposed by Arikan [1] for binary memoryless symmetric channels and generalized for Galois fields [2] and arbitrary q-ary alphabets [3]. The idea of polar codes is also extended to lossless and lossy source coding and some multiterminal problems [4][5][6].

We consider channel coding with polar codes for asymmetric memoryless channels and lossy source coding for nonuniform sources and/or asymmetric distortion measures. In these asymmetric settings, the optimal symbol distribution of codewords to achieve the Shannon bound is not always uniform.

In known polar coding schemes for asymmetric settings, codewords with a nonuniform symbol distribution are generated based on Gallager's scheme [7, p. 208], which uses nonlinear mapping of symbols illustrated by the following example. Consider channel coding of an asymmetric channel such that the optimal input distribution is $(P_X(0), P_X(1)) = (2/3, 1/3)$ with alphabet $\mathcal{X} = \{0, 1\}$. This input distribution can be realized by considering a ternary polar code with an extended alphabet $\mathcal{X}' = \{a, b, c\}$. Mapping symbols $a, b \in \mathcal{X}'$ to $0 \in \mathcal{X}$ and $c \in \mathcal{X}'$ to $1 \in \mathcal{X}$ in codewords, we obtain codewords on $\mathcal{X}$ with the desired distribution. Although this technique is simple and applicable widely, a large size of extended alphabet is required if the optimal distribution $P_X(\cdot)$ cannot be approximated by simple rational numbers. In such cases, the complexity of decoding increases considerably. (See Section V for a detailed discussion on the complexity of Gallager's scheme.)

To overcome the defect of Gallager's scheme, we need to generate the given symbol distribution $P_X(\cdot)$ of codewords without any extended alphabet. A key idea to generate a desired distribution can be found in the lossless compression by polar codes [6]. In this setting an original message $X_1^n = (X_1, X_2, \ldots, X_n)$ with a nonuniform distribution is transformed to $U_1^n = X_1^n G_n$ by the generator matrix G_n of polar codes. It is shown that the elements of U_1^n polarize into two groups, $\mathcal{F}$ and $\mathcal{F}^c$. For each $i \in \mathcal{F}$, U_i is almost uniformly distributed and independent of the leading sequence $U_1^{i-1} = (U_1, U_2, \ldots, U_{i-1})$ and, for each $i \in \mathcal{F}^c$, U_i is determined from U_1^{i-1} almost surely.

Now we apply this technique to channel coding. The result on the lossless coding implies that, when we have a uniform source, we can obtain a nonuniform input for a given channel in the following way: (a) choose a value of U_i uniformly for each $i \in \mathcal{F}$, (b) determine U_i for each $i \in \mathcal{F}^c$ appropriately from U_1^{i-1} and (c) transform U_1^n to X_1^n by $X_1^n = U_1^n G_n^{-1} = U_1^n G_n$. In the case of channel coding with channel input X and channel output Y, U_i for each $i \in \mathcal{F}$ polarizes further into $\mathcal{I} \subset \mathcal{F}$ and $\mathcal{F} \setminus \mathcal{I}$, where this polarization corresponds to lossless source coding with side information [8]. Here $\mathcal{I}$ is the set of indices i such that U_i is almost independent of U_1^{i-1} but can be determined uniquely given U_1^{i-1} and channel output Y_1^n almost surely. $\mathcal{F} \setminus \mathcal{I}$ is the set of indices i such that U_i is almost independent of both U_1^{i-1} and Y_1^n. By assigning a message to random variables U_i for $i \in \mathcal{I}$ we can send it with small decoding error probability, that is, $\mathcal{I} \subset \mathcal{F}$ can be used as an information set. The relation between lossless source coding and channel coding is depicted in Fig. 1.

This idea can also be applied to lossy source coding. In this case Y_1^n and X_1^n correspond to a source sequence and a reproduction sequence, respectively. In lossy source coding, U_i which is almost random given (U_1^{i-1}, Y_1^n) does not need to be sent because such U_i does not affect the joint distribution of (X_1^n, Y_1^n) even if U_i is determined independently of Y_1^n. Furthermore, in the asymmetric case, U_i which is almost deterministic given U_1^{i-1} also does not need to be sent because it can be reproduced in the same way as in lossless coding. As a result, by sending U_i only for $i \in \mathcal{I}$, we can recover X_1^n within a given distortion.

Note that recently Sutter et al. [9] have considered a channel

The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Cambridge, USA, July 2012.

The authors are with the Department of Complexity Science and Engineering, The University of Tokyo, Kashiwa-shi Chiba 277–8561, Japan, (e-mail: honda@it.k.u-tokyo.ac.jp; Hirosuke@ieee.org).

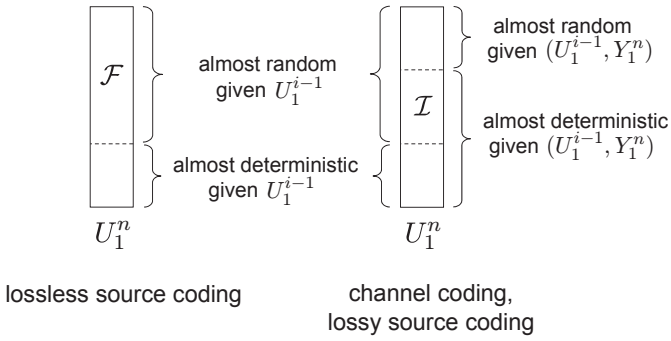

Fig. 1. Polarization of $U_1^n = X_1^n G_n$ for nonuniform X_1^n with/without side information Y_1^n.

coding scheme for asymmetric settings based on lossless coding independently of our work. However, their scheme uses a concatenated code of two polar codes and the code construction is not simple. Furthermore, the decoding error probability of their scheme is approximately $\mathrm{O}(2^{-n^{1/4}})$ since their scheme requires polarization for both the inner code and the outer code, whereas we show that the decoding error probability of our scheme is approximately $\mathrm{O}(2^{-n^{1/2}})$ by using a single polar code.

This paper is organized as follows. In Section II, we introduce polar codes for lossless compression with side information considered in [8] and derive a polarization phenomenon required for our setting. In Section III, we propose a new polar coding scheme for asymmetric channels and show that it can achieve the channel capacity asymptotically. We show in Section IV that the same idea can also be applied to lossy source coding with nonuniform sources and/or asymmetric distortion measures. In Section V, we compare the proposed scheme with Gallager's scheme in view of coding rates, complexities and decoding error probabilities. We give proofs of theorems in the appendices.

II. Polarization for Asymmetric Settings

Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a pair of random variables. In the case of lossless compression, X and Y correspond to a source and side information, respectively. For simplicity we assume that $\mathcal{X}$ is binary, i.e., $\mathcal{X} = \{0, 1\} = \mathrm{GF}(2)$, but $\mathcal{Y}$ is an arbitrary finite set. We note that the following results can be extended to nonbinary cases, i.e., $|\mathcal{X}| \geq 3$ in the same way as the case of symmetric channels (see [2][3] and [8, Section VI]). The addition on GF(2) is denoted by operator $\oplus$.

Let $X_1^n = (X_1, X_2, \ldots, X_n)$ and $Y_1^n = (Y_1, Y_2, \ldots, Y_n)$ denote i.i.d. copies of X and Y, respectively. For $n = 2^k$, the generator matrix of polar codes[1] is given by $G_n = G^{\otimes k}$ where $G = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ and $\otimes$ denotes the Kronecker power. U_1^n is defined as $U_1^n \equiv X_1^n G_n^{-1} = X_1^n G_n$. Let U_i^j, $i < j$, stand for subvector $(U_i, U_{i+1}, \ldots, U_j)$ of U_1^n. Similarly, let $U_{\mathcal{A}}$, $\mathcal{A} \subset \{1, 2, \ldots, n\}$, represent subvector $\{U_i\}_{i \in \mathcal{A}}$.

In the case of channel coding for symmetric binary-input discrete memoryless channels W, Bhattacharyya parameter $Z_{\mathrm{B}}(W)$, which is defined by

$$Z_{\mathrm{B}}(W) \equiv \sum_y \sqrt{W(y|0)W(y|1)}\,, \tag{1}$$

is used to evaluate the error probability. For the case of source coding with side information, this parameter is extended to $Z(X|Y)$ defined as

$$\begin{aligned} Z(X|Y) &\equiv 2\sum_y P_Y(y)\sqrt{P_{X|Y}(0|y)P_{X|Y}(1|y)} \\ &= 2\sum_y \sqrt{P_{X,Y}(0,y)P_{X,Y}(1,y)}\,. \end{aligned} \tag{2}$$

Note that $Z(X|Y)$ coincides with the Bhattacharyya parameter $Z_{\mathrm{B}}(P_{Y|X})$ when X is uniformly distributed. The parameter $Z(X|Y)$ is related to conditional entropy $H(X|Y)$ by the following proposition.

Proposition 1 ([8, Proposition 2])**.**

$$(Z(X|Y))^2 \leq H(X|Y)\,, \tag{3}$$

$$H(X|Y) \leq \log(1 + Z(X|Y)) \leq Z(X|Y)\,. \tag{4}$$

Now we give the main result of this section on the polarization for asymmetric cases.

Theorem 1. *For any $\beta < 1/2$, i.i.d. random variables (X, Y) and $U_1^n = X_1^n G_n$,*

$$\lim_{n\to\infty} \frac{1}{n}\left|\left\{i : Z(U_i|U_1^{i-1}, Y_1^n) \leq 2^{-n^\beta} \text{ and } Z(U_i|U_1^{i-1}) \geq 1 - 2^{-n^\beta}\right\}\right| = I(X;Y)\,, \tag{5}$$

$$\lim_{n\to\infty} \frac{1}{n}\left|\left\{i : Z(U_i|U_1^{i-1}, Y_1^n) \geq 1 - 2^{-n^\beta} \text{ or } Z(U_i|U_1^{i-1}) \leq 2^{-n^\beta}\right\}\right| = 1 - I(X;Y)\,. \tag{6}$$

We prove this theorem in Appendix B using a technique similar to that for lossless coding by polar codes in [6][8]. In [8], a recursive formula is derived for $Z(U_i|U_1^{i-1}, Y_1^n)$ and the polarization of $Z(U_i|U_1^{i-1}, Y_1^n)$ is shown from the fact that the formula has the same form as the symmetric case. On the other hand, in [6], the asymptotic optimality is derived by reducing the source coding problem to a channel coding one. In this paper we apply this reduction technique to parameter $Z(U_i|U_1^{i-1}, Y_1^n)$ and show that $Z(U_i|U_1^{i-1}, Y_1^n)$ is equal to a Bhattacharyya parameter $Z_{\mathrm{B}}(\tilde{W})$ for some symmetric channel $\tilde{W}$. By this representation we can apply known results on the symmetric settings directly to our asymmetric settings.

Now we review the polarization for symmetric channels. Let $\tilde{W}$ be a *symmetric* binary-input discrete memoryless channel with transition probability $\tilde{W}(\tilde{y}|\tilde{x})$ for $\tilde{x} \in \mathcal{X} = \{0, 1\}$ and $\tilde{y} \in \tilde{\mathcal{Y}}$. $(\tilde{X}, \tilde{Y})$ is a pair of random variables with distribution $P_{\tilde{X}\tilde{Y}}(\tilde{x}, \tilde{y}) = P_{\tilde{X}}(\tilde{x})\tilde{W}(\tilde{y}|\tilde{x})$, where $P_{\tilde{X}}$ is the uniform distribution on $\mathcal{X}$. $(\tilde{X}_1^n, \tilde{Y}_1^n)$ is a sequence of n i.i.d. copies of

[1] More precisely, the generator matrix of a polar code is a submatrix of G_n. But for simplicity we call G_n the generator matrix of polar codes in this paper. Note also that the generator matrix is sometimes defined as $G_n = B_n G^{\otimes k}$ instead of $G_n = G^{\otimes k}$ for the bit-reversal matrix B_n, and both definitions are essentially equivalent (see [1, Section VII-C]).

$(\tilde{X}, \tilde{Y})$. $\tilde{U}_1^n$ is defined by $\tilde{U}_1^n = \tilde{X}_1^n G_n^{-1} = \tilde{X}_1^n G_n$. The ith subchannel for a polar code is given by

$$\tilde{W}_i^{(n)}(\tilde{u}_1^{i-1}, \tilde{y}_1^n | \tilde{u}_i) = P_{\tilde{U}_1^{i-1}, \tilde{Y}_1^n | \tilde{U}_i}(\tilde{u}_1^{i-1}, \tilde{y}_1^n | \tilde{u}_i)\,. \tag{7}$$

The symmetric capacity of channel $\tilde{W}$ is given by $I(\tilde{W}) = I(\tilde{X}; \tilde{Y})$. The polarization for this symmetric channel is described as follows.

Proposition 2 ([6, Theorems 2.11 and 3.15])**.** *For any symmetric binary-input discrete memoryless channel $\tilde{W}$, $\beta < 1/2$ and $\tilde{W}_i^{(n)}$ defined by* (7)*,*

$$\lim_{n\to\infty} \frac{1}{n} \left| \left\{ i : Z_{\mathrm{B}}(\tilde{W}_i^{(n)}) \le 2^{-n^\beta} \right\} \right| = I(\tilde{W})\,, \tag{8}$$

$$\lim_{n\to\infty} \frac{1}{n} \left| \left\{ i : Z_{\mathrm{B}}(\tilde{W}_i^{(n)}) \ge 1 - 2^{-n^\beta} \right\} \right| = 1 - I(\tilde{W})\,. \tag{9}$$

Recall that X_1^n and Y_1^n are i.i.d. copies of $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, respectively, and $U_1^n = X_1^n G_n$. The following theorem enables us to apply known results on symmetric channels including Proposition 2 to our asymmetric setting.

Theorem 2. *Let $\tilde{\mathcal{Y}} = \{0,1\} \times \mathcal{Y}$ and $\tilde{Y}_1^n = (\tilde{X}_1^n \oplus X_1^n, Y_1^n)$ where (X_1^n, Y_1^n) is independent of $\tilde{X}_1^n$. Then, for $\tilde{W}_i^{(n)}$ defined by* (7)*,*

$$P_{U_1^i, Y_1^n}(u_1^i, y_1^n) = 2^{n-1} \tilde{W}_i^{(n)}(u_1^{i-1}, (0^n, y_1^n) | u_i) \tag{10}$$

and

$$Z(U_i | U_1^{i-1}, Y_1^n) = Z_{\mathrm{B}}(\tilde{W}_i^{(n)})\,. \tag{11}$$

We prove this theorem in Appendix A.

III. Polar Codes for Asymmetric Channels

In this section we propose a new polar coding scheme which can achieve the capacity for asymmetric memoryless channels.

A. Code Construction

Assume that an information set $\mathcal{I} \subset \{1, 2, \ldots, n\}$ and a frozen set $\mathcal{I}^c = \{1, 2, \ldots, n\} \setminus \mathcal{I}$ are fixed for a given channel W. We use bits $u_{\mathcal{I}} = \{u_i\}_{i \in \mathcal{I}}$ to send a message.

In the case of symmetric channels, the values of frozen bits $u_{\mathcal{I}^c}$ are chosen randomly with the uniform distribution on $\{0,1\}$ in the code construction but they are fixed when the code is used. In our scheme, the frozen bits $u_{\mathcal{I}^c}$ are deterministic but dependent on the value of previous bits u_1^{i-1}. Furthermore, unlike the symmetric case, we choose the value of u_i given u_1^{i-1} *not uniformly* in the code construction.

Let $\mathcal{L}_i$ be the family of functions $\lambda_i : \{0,1\}^{i-1} \to \{0,1\}$. Now we consider a polar code with frozen set $\mathcal{I}^c$ and maps $\lambda_{\mathcal{I}^c} \equiv \{\lambda_i\}_{i \in \mathcal{I}^c}$. The maps $\lambda_{\mathcal{I}^c}$ are used to determine the frozen bits and are shared between the encoder and the decoder.

Let $M_1^{|\mathcal{I}|}$ denote a message uniformly distributed on $\{0,1\}^{|\mathcal{I}|}$. The encoder determines a codeword from a realization $m_1^{|\mathcal{I}|}$ of $M_1^{|\mathcal{I}|}$ in the following way. First, the encoder determines the information bits by $u_{\mathcal{I}} = m_1^{|\mathcal{I}|}$. Next, for the frozen bits $\mathcal{I}^c$, the encoder determines the value u_i, $i \in \mathcal{I}^c$, in the ascending order by $u_i = \lambda_i(u_1^{i-1})$. We represent the resulting sequence of u_i by $u_1^n(m_1^{|\mathcal{I}|}, \lambda_{\mathcal{I}^c})$. Third, the encoder sends the codeword $x_1^n = u_1^n G_n = u_1^n(m_1^{|\mathcal{I}|}, \lambda_{\mathcal{I}^c})\, G_n$ with code length n. Thus the coding rate is given by $R = |\mathcal{I}|/n$.

The decoder receives a sequence y_1^n according to the channel transition probability $W^n(y_1^n | x_1^n)$. The decoder estimates u_1^n by $\hat{u}_1^n = \hat{u}_1^n(y_1^n, \lambda_{\mathcal{I}^c})$ as follows:

$$\hat{u}_i = \begin{cases} \underset{u}{\operatorname{argmax}}\, P_{U_i | U_1^{i-1}, Y_1^n}(u | \hat{u}_1^{i-1}, y_1^n) & i \in \mathcal{I}, \\ \lambda_i(\hat{u}_1^{i-1}) & i \in \mathcal{I}^c. \end{cases} \tag{12}$$

The decoding is successful if $\hat{u}_{\mathcal{I}} = u_{\mathcal{I}}$ which means $\hat{u}_1^n = u_1^n$. The average decoding error probability over the uniform message $M_1^{|\mathcal{I}|}$ is denoted by $P_{\mathrm{e}}(\lambda_{\mathcal{I}^c})$.

Now consider the choice of the map $\lambda_{\mathcal{I}^c}$. Let $\Lambda_{\mathcal{I}^c} \equiv \{\Lambda_i \in \mathcal{L}_i\}_{i \in \mathcal{I}^c}$ be random variables which are independent of each other and of (X_1^n, Y_1^n), and satisfy

$$P_{\Lambda_i}[\Lambda_i(u_1^{i-1}) = 1] = P_{U_i | U_1^{i-1}}(1 | u_1^{i-1}) \tag{13}$$

for all $u_1^{i-1} \in \{0,1\}^{i-1}$. Practically, we can realize this randomized map by using pseudo random numbers shared between the encoder and the decoder as follows.

$$u_i = \begin{cases} 0 & \text{with probability } P_{U_i | U_1^{i-1}}(0 | u_1^{i-1})\,, \\ 1 & \text{with probability } P_{U_i | U_1^{i-1}}(1 | u_1^{i-1})\,. \end{cases} \tag{14}$$

The idea of this randomized algorithm comes from the polar coding for lossy compression for symmetric sources [5][6], where this technique is called *randomized rounding*. As in the case of the lossy coding for symmetric sources, the randomization makes the theoretical analysis much easier in our setting.

From Theorem 1 there exists a subset $\mathcal{I}$ of $\{1, \cdots, n\}$ such that $|\mathcal{I}| = nR$,

$$Z(U_i | U_1^{i-1}, Y_1^n) \le 2^{-n^\beta} \ \text{ and } \ Z(U_i | U_1^{i-1}) \ge 1 - 2^{-n^\beta} \tag{15}$$

for all $i \in \mathcal{I}$ if $R < I(X;Y)$, $\beta < 1/2$, and n is sufficiently large. For this $\mathcal{I}$ the following theorem holds.

Theorem 3. *Let $M_1^{|\mathcal{I}|}$ be a message chosen uniformly from $\{0,1\}^{|\mathcal{I}|}$ and $\mathcal{I} \subset \{1, \cdots, n\}$ be a set satisfying* (15)*. Then the expectation of the decoding error probability over the maps $\Lambda_{\mathcal{I}^c}$ satisfies $\mathrm{E}_{\Lambda_{\mathcal{I}^c}}[P_{\mathrm{e}}(\Lambda_{\mathcal{I}^c})] = \mathrm{O}(2^{-n^{\beta'}})$ for any $\beta' < \beta < 1/2$. Consequently, there exists a deterministic map $\lambda_{\mathcal{I}^c} = \{\lambda_i \in \mathcal{L}_i\}_{i \in \mathcal{I}^c}$ such that $P_{\mathrm{e}}(\lambda_{\mathcal{I}^c}) = \mathrm{O}(2^{-n^{\beta'}})$.*

The proof of this theorem is given in Appendix C.

B. Implementation

In the construction of the proposed coding scheme, information set $\mathcal{I}$ has to be chosen from $\{1, \ldots, n\}$ so that $Z(U_i | U_1^{i-1}, Y_1^n)$ is small and $Z(U_i | U_1^{i-1})$ is large for every $i \in \mathcal{I}$. From Theorem 2 these parameters can be represented as Bhattacharyya parameters for symmetric channels and the approximation technique in [10] for symmetric cases can be applied.

In the encoding of the proposed polar coding scheme, probability $P_{U_i | U_1^{i-1}}(u | u_1^{i-1})$ in (14) has to be computed. Similarly, in the decoding, we need to compute $P_{U_i | U_1^{i-1}, Y_1^n}(u | \hat{u}_1^{i-1}, y_1^n)$

in (12) and $P_{U_i|U_1^{i-1}}(u|u_1^{i-1})$ in (14). From (10) in Theorem 2, we can represent the ratio of the posterior probability by

$$\frac{P_{U_i|U_1^{i-1},Y_1^n}(1|\hat{u}_1^{i-1},y_1^n)}{P_{U_i|U_1^{i-1},Y_1^n}(0|\hat{u}_1^{i-1},y_1^n)} = \frac{P_{U_1^i,Y_1^n}((\hat{u}_1^{i-1},1),y_1^n)}{P_{U_1^i,Y_1^n}((\hat{u}_1^{i-1},0),y_1^n)} = \frac{\tilde{W}_i^{(n)}((0^n,y_1^n),\hat{u}_1^{i-1}|1)}{\tilde{W}_i^{(n)}((0^n,y_1^n),\hat{u}_1^{i-1}|0)}. \quad (16)$$

Since the RHS of (16) can be computed with complexity $\mathrm{O}(n\log n)$ by the decoding technique of polar codes for symmetric channel [1], we can also compute $P_{U_i|U_1^{i-1},Y_1^n}(u|\hat{u}_1^{i-1},y_1^n)$ with complexity $\mathrm{O}(n\log n)$. We can compute $P_{U_i|U_1^{i-1}}(u|u_1^{i-1})$ similarly by letting Y be a constant random variable.

IV. Application to Lossy Source Coding

In this section we consider polar coding for nonuniform sources and/or asymmetric distortion measures.

For information source $Y \in \mathcal{Y}$ and distortion measure $d : \mathcal{Y}\times\{0,1\} \to [0,+\infty)$, the rate-distortion function is given by

$$R(D) = \min_{X':\mathrm{E}_{XY}[d(Y,X')]\le D} I(X';Y). \quad (17)$$

The random variable achieving this minimum is denoted by X in the following.

Now we construct a polar code for the lossy coding problem. Assume that an information set $\mathcal{I} \subset \{1,\dots,n\}$ and a frozen set $\mathcal{I}^c = \{1,\dots,n\}\setminus\mathcal{I}$ are given and satisfy $|\mathcal{I}| = nR$ and

$$Z(U_i|U_1^{i-1},Y_1^n) \ge 1-2^{-n^\beta} \text{ or } Z(U_i|U_1^{i-1}) \le 2^{-n^\beta} \quad (18)$$

for all $i \in \mathcal{I}^c$. From Theorem 1, such $\mathcal{I}$ exists if $R > I(X;Y) = R(D)$, $\beta < 1/2$ and n is sufficiently large.

As in the case of channel coding, let $\mathcal{L}_i$ be the family of functions $\lambda_i : \{0,1\}^{i-1} \to \{0,1\}$ and assume that $\lambda_{\mathcal{I}^c} \in \prod_{i\in\mathcal{I}^c}\mathcal{L}_i$ is shared between the encoder and the decoder. In the proposed scheme, the encoder determines $u_1^n = u_1^n(\lambda_{\mathcal{I}^c},y_1^n)$ from a given source sequence y_1^n by

$$u_i = \begin{cases} 0 & \text{with probability } P_{U_i|U_1^{i-1},Y_1^n}(0|u_1^{i-1},y_1^n) \\ 1 & \text{with probability } P_{U_i|U_1^{i-1},Y_1^n}(1|u_1^{i-1},y_1^n) \end{cases} \quad (19)$$

for $i \in \mathcal{I}$ and

$$u_i = \lambda_i(u_1^{i-1}) \quad (20)$$

for $i \in \mathcal{I}^c$. The encoder sends $u_{\mathcal{I}}$ to the decoder. The decoder determines $u_{\mathcal{I}^c}$ by $u_i = \lambda_i(u_1^{i-1})$ and outputs reproduction sequence $x_1^n = u_1^n G_n$. Then, the coding rate is given by $R = |\mathcal{I}|/n$. We define the average distortion by

$$D_n(\lambda_{\mathcal{I}^c}) \equiv \frac{1}{n}\mathrm{E}_{Y_1^n}\left[\mathrm{E}[d^n(Y_1^n,u_1^n(\lambda_{\mathcal{I}^c},Y_1^n)\,G_n)]\right] \quad (21)$$

where $d^n(y_1^n,x_1^n) \equiv \sum_{i=1}^n d(y_i,x_i)$ and the inner expectation is taken over randomization in (19).

As in the case of channel coding, we consider the expectation of $D_n(\Lambda_{\mathcal{I}^c})$ for random variable $\Lambda_{\mathcal{I}^c}$ such that $P_{\Lambda_i}[\Lambda_i(u_1^{i-1}) = 1] = P_{U_i|U_1^{i-1}}(1|u_1^{i-1})$ for all $u_1^{i-1} \in \{0,1\}^{i-1}$.

Theorem 4. *Let $\mathcal{I} \subset \{1,\dots,n\}$ be a set satisfying* (18)*. Then the expectation of the average distortion $D_n(\Lambda_{\mathcal{I}^c})$ over the maps $\Lambda_{\mathcal{I}^c}$ satisfies $\mathrm{E}_{\Lambda_{\mathcal{I}^c}}[D_n(\Lambda_{\mathcal{I}^c})] = D + \mathrm{O}(2^{-n^{\beta'}})$ for any $R > R(D)$ and $\beta' < \beta < 1/2$. Consequently, there exists a deterministic map $\lambda_{\mathcal{I}^c} = \{\lambda_i \in \mathcal{L}_i\}_{i\in\mathcal{I}^c}$ such that $D_n(\lambda_{\mathcal{I}^c}) = D + \mathrm{O}(2^{-n^{\beta'}})$.*

The proof follows the same line as that of Theorem 3 and is given in Appendix D.

Remark 1. In the lossy coding for symmetric setting [6], $u_{\mathcal{I}^c}$ is determined beforehand uniformly from $\{0,1\}^{|\mathcal{I}^c|}$ and the randomized map $\Lambda_{\mathcal{I}^c}$ for the frozen set is not required. In our setting, the achievability of the rate-distortion function can be proved in the similar way as [6] for a simplified rule such that for $i \in \mathcal{I}^c$

$$u_i = \begin{cases} \bar{u}_i, & \text{if } Z(U_i|U_1^{i-1},Y_1^n) \ge 1-2^{-n^\beta}, \\ \operatorname{argmax}_u P_{U_i|U_1^{i-1}}(u|u_1^{i-1}), & \text{if } Z(U_i|U_1^{i-1}) \le 2^{-n^\beta}, \end{cases} \quad (22)$$

instead of (20), where $\bar{u}_i$ is determined beforehand uniformly from $\{0,1\}$. However, since this rule makes the proof of Theorem 4 a little longer, we use the map $\Lambda_{\mathcal{I}^c}$ for simplicity although $\Lambda_{\mathcal{I}^c}$ has to be shared between the encoder and the decoder.

V. Comparison with Gallager's Scheme

In this section we compare the proposed scheme with Gallager's scheme using the alphabet extension [7]. We mainly consider polar codes with generator matrix $G_n = G^{\otimes k}$ for a 2×2 matrix G over $\mathrm{GF}(q)$ used in [2][3] to implement Gallager's scheme and later discuss other implementations. In Gallager's scheme we only treat the case that block length n is taken sufficiently large for a *fixed* size q of the extended alphabet although one may be interested in the case that q increases with the block length. It is because, in Gallager's scheme, the original channel is transformed into a q-input $|\mathcal{Y}|$-output channel, which has different characteristics depending on q. Since the asymptotic decoding error probability $\mathrm{O}(2^{-n^\beta})$ in, e.g., Theorem 3 or [11, Theorem 1] of polar codes is derived for a fixed channel and the dependency on the channel is neglected, it is very difficult to analyze the asymptotic performance of Gallager's scheme with increasing q. Hence, we do not consider such cases in this paper.

A. Coding Rate and Complexity

First we consider the coding rate of Gallager's scheme. The achievable rate by input distribution $\boldsymbol{p} = (p_0,p_1) = (P_X(0),P_X(1))$ is given by

$$I(\boldsymbol{p}) = \sum_{x,y} P_X(x)W(y|x)\log\frac{P_X(x)W(y|x)}{P_X(x)\sum_{x'}P_X(x')W(y|x')} \quad (23)$$

with the optimal input distribution

$$\boldsymbol{p}^* = \operatorname*{argmax}_{\boldsymbol{p}:\sum_i p_i=1} I(\boldsymbol{p}). \quad (24)$$

From the smoothness of $I(\boldsymbol{p})$, the objective function is approximated by

$$\begin{aligned} I(\boldsymbol{p}) &= I(\boldsymbol{p}^*) + (\boldsymbol{p}-\boldsymbol{p}^*)\mathbf{H}(I(\boldsymbol{p}))|_{\boldsymbol{p}=\boldsymbol{p}^*}(\boldsymbol{p}-\boldsymbol{p}^*)^\top \\ &\quad + \mathrm{o}(\|\boldsymbol{p}-\boldsymbol{p}^*\|^2) \\ &= C(W) + (\boldsymbol{p}-\boldsymbol{p}^*)\mathbf{H}(I(\boldsymbol{p}))|_{\boldsymbol{p}=\boldsymbol{p}^*}(\boldsymbol{p}-\boldsymbol{p}^*)^\top \\ &\quad + \mathrm{o}(\|\boldsymbol{p}-\boldsymbol{p}^*\|^2) \end{aligned} \tag{25}$$

in the neighborhood of $\boldsymbol{p}^*$ where $\mathbf{H}(I)$ is the Hessian of I and $\top$ denotes the transpose[2]. Gallager's scheme can approximate the optimal distribution $\boldsymbol{p}^*$ by a distribution $\boldsymbol{p}$ such that $\|\boldsymbol{p}-\boldsymbol{p}^*\| = \mathrm{O}(1/q)$ and the gap of the achievable rate is expressed as $\Delta I = I(\boldsymbol{p}^*) - I(\boldsymbol{p}) = \mathrm{O}(\|\boldsymbol{p}-\boldsymbol{p}^*\|^2) = \mathrm{O}(1/q^2)$. On the other hand, the successive cancellation of polar codes requires convolutions of q-ary symbols, a naive implementation of which has complexity $\mathrm{O}(q^2)$ and it can be reduced to $\mathrm{O}(q\log q)$ by the technique of fast Fourier transformation when q is a power of prime as mentioned in [13, Section III-B], although it has not been shown that they are the best possible. As a result, Gallager's scheme currently requires complexity at least $\mathrm{O}((\Delta I)^{-1/2} n\log n)$. In contrast, our scheme always bounds the complexity by $\mathrm{O}(n\log n)$ independent of the coding rate.

We can also consider the coding problem of an AWGN channel with an average power constraint. In this problem the optimal input distribution to achieve the capacity $C(W) = (1/2)\log(1+\mathrm{SNR})$ is a normal distribution, which has to be approximated by a discrete distribution in practice. As discussed in [9], the gap of the capacity of Gallager's scheme is given by $\mathrm{O}(1/q)$ [13], but that of the proposed scheme is exponentially small in q [14]. Therefore, the proposed scheme can achieve the same coding rate as that of Gallager's scheme by a logarithmically small alphabet size q.

It is worth noting that the optimal input distribution of an AWGN channel has a discrete finite support if the peak power of the channel input is constrained in addition to the average power [15]. Furthermore, many continuous-output channels have discrete optimal input distributions under the average power constraint and/or the peak power constraints [16]. In such cases, the proposed scheme can realize the optimal input distribution without alphabet extension.

B. Decoding Error Probability

Next we compare the decoding error probabilities of these schemes. In the second-order analysis for binary polar codes [17][18][19], it is shown that the decoding error probability can be expressed as

$$P_{\mathrm{e}} = 2^{-2^{\frac{k}{2}+\frac{\sqrt{k}}{2}Q^{-1}\left(\frac{R}{C(W)}\right)+\mathrm{o}(\sqrt{k})}} \tag{26}$$

for any symmetric channel W, where Q^{-1} is the inverse of the error function $Q(t) = \int_t^{+\infty}(2\pi)^{-1/2}\exp(-s^2/2)\mathrm{d}s$. This result can be extended to polar codes with a prime size q of a channel input (see [20, Section VII]) and the error probability can be written in the same form as (26), say

$$P_{\mathrm{e,Gallager}} = 2^{-2^{\frac{k}{2}+\frac{\sqrt{k}}{2}Q^{-1}\left(\frac{R}{I(\boldsymbol{p})}\right)+\mathrm{o}(\sqrt{k})}}, \tag{27}$$

for Gallager's scheme. On the other hand in the proposed scheme, we have

$$\begin{aligned} &\lim_{n\to\infty}\frac{1}{n}\left|\left\{ i : Z(U_i|U_1^{i-1}) \ge 1-2^{-2^{\frac{k}{2}+\frac{\sqrt{k}}{2}Q^{-1}\left(\frac{R_2}{H(X)}\right)+\mathrm{o}(\sqrt{k})}} \right.\right. \\ &\qquad \left.\left. Z(U_i|U_1^{i-1},Y_1^n) \le 2^{-2^{\frac{k}{2}+\frac{\sqrt{k}}{2}Q^{-1}\left(\frac{1-R_1}{1-H(X|Y)}\right)+\mathrm{o}(\sqrt{k})}} \right\}\right| \\ &= R_2 - R_1\,, \end{aligned} \tag{28}$$

for any $R_1 > H(X|Y)$ and $R_2 < H(X)$ by applying the second-order analysis to (5) in Theorem 1. This result induces a polar code with coding rate $R = R_2 - R_1$ with error probability

$$\begin{aligned} P_{\mathrm{e,propose}} &= 2^{-2^{\frac{k}{2}+\frac{\sqrt{k}}{2}Q^{-1}\left(\frac{1-R_1}{1-H(X|Y)}\right)+\mathrm{o}(\sqrt{k})}} \\ &\quad + 2^{-2^{\frac{k}{2}+\frac{\sqrt{k}}{2}Q^{-1}\left(\frac{R_2}{H(X)}\right)+\mathrm{o}(\sqrt{k})}} \\ &= 2^{-2^{\frac{k}{2}+\frac{\sqrt{k}}{2}Q^{-1}\left(\max\left\{\frac{1-R_1}{1-H(X|Y)},\frac{R_2}{H(X)}\right\}\right)+\mathrm{o}(\sqrt{k})}}, \end{aligned} \tag{29}$$

which is optimized as

$$P_{\mathrm{e,propose}} = 2^{-2^{\frac{k}{2}+\frac{\sqrt{k}}{2}Q^{-1}\left(\frac{1+R}{1+C(W)}\right)+\mathrm{o}(\sqrt{k})}} \tag{30}$$

by setting $R_1 = (H(X) - R(1-H(X|Y)))/(1+C(W))$ and $R_2 = H(X)(1+R)/(1+C(W))$. Note that

$$\frac{1+R}{1+C(W)} > \frac{R}{C(W)} \tag{31}$$

always holds for $R < C(W) = I(\boldsymbol{p}^*)$. Therefore, in view of (27) and (30), the decoding error probability of the proposed scheme is asymptotically worse than Gallager's scheme if it can exactly realize the optimal input distribution $\boldsymbol{p}^*$. However, in the case that the input distribution $\boldsymbol{p}$ of Gallager's scheme is different from $\boldsymbol{p}^*$, the proposed scheme has a better second-order exponent if

$$\begin{aligned} &\frac{1+R}{1+C(W)} < \frac{R}{I(\boldsymbol{p})} \\ &\iff R > I(\boldsymbol{p}) - \tfrac{C(W)-\Delta I}{1+\Delta I}\Delta I \approx I(\boldsymbol{p}) - C(W)\Delta I \end{aligned} \tag{32}$$

where $\Delta I = C(W) - I(\boldsymbol{p})$. The analysis of the error probability is summarized as follows: (a) if $C(W) = I(\boldsymbol{p}^*) > R > I(\boldsymbol{p})$ then only the proposed scheme can achieve any small error probability, (b) if $I(\boldsymbol{p}) > R \gtrsim I(\boldsymbol{p}) - C(W)\Delta I$ then both schemes achieve any small error probability but the proposed scheme has a better second-order error exponent, and (c) if $I(\boldsymbol{p}) - C(W)\Delta I \gtrsim R$ then Gallager's scheme has a better second-order error exponent.

[2]Refer, e.g., [12, Theorem 4] and its proof for detail, where the second-order approximation of the objective function is considered to derive a necessary condition for an optimal solution of a constrained optimization problem.

C. Extension to General Kernels

In the above analyses we have mainly considered polar codes with generator matrix $G^{\otimes k}$ for 2×2 matrices G over $\mathrm{GF}(2)$ in the proposed scheme and $\mathrm{GF}(q)$ in Gallager's scheme. In this section, we discuss generalization of the kernel G to $l \times l$ matrices over $\mathrm{GF}(\tilde{q})$ by regarding $\log_2 \tilde{q}$ symbols in the proposed scheme (or $\log_q \tilde{q}$ symbols in Gallager's scheme) as one symbol over $\mathrm{GF}(\tilde{q})$.

It is shown in [2][20][21] that the decoding error probability can be expressed as $\mathrm{O}(2^{-l^{kE(G)}})$ for some first-order error exponent $E(G) \in (0,1)$, which can be increased when we use large l and $\tilde{q}$. Here note that $E(G)$ can be maximized by $l \times l$ Reed-Solomon matrix over $\mathrm{GF}(l)$ with exponent $E(G_{\mathrm{RS}}) = \log(l!)/(l \log l)$ when the size of the matrix l is fixed [20]. This means that, although a larger $\tilde{q}$ increases the complexity, $\tilde{q}$ larger than l does not improve the error exponent.

On the other hand in Gallager's scheme, $\tilde{q} \geq q$ is required where q depends on the coding rate to be achieved. Thus we need to use $l \geq q$ to take full advantage of the extended alphabet size q of Gallager's scheme, a naive implementation of which seems to require complexity $\mathrm{O}(q^l n \log n) \geq \mathrm{O}(q^q n \log n)$ (even if the technique of fast Fourier transformation is used). Then we conclude that the proposed scheme can achieve a better tradeoff between the (first-order) error exponent and the complexity than Gallager's scheme unless exponential complexity in q is allowed, since we can design a kernel G without any constraint on the alphabet size $\tilde{q}$ in the proposed scheme.

Note that a technique of multiple access channel (MAC) coding [22] can also be applied to Gallager's scheme. In this technique, $m = \log q$ virtual users transmit $x[1], x[2], \cdots, x[m] \in \{0,1\}$ via a single channel with transition probability $W_{\mathrm{MAC}}(y|\{x[j]\}_{j=1,2,\cdots,m})$. By letting

$$W_{\mathrm{MAC}}(y|\{x[j]\}_j) = \begin{cases} W(y|0), & \text{if } \{x[j]\}_j \subset S, \\ W(y|1), & \text{otherwise} \end{cases} \tag{33}$$

for some $S \subset \{0,1\}^m$ such that $|S| = 2^m P_X(0)$, we can express the coding problem of asymmetric channels as a MAC coding with sum-rate $I(\{X[j]\}_j; Y)$ for the uniform input $\{X[j]\}_j \in \{0,1\}^m$. This uniform sum rate can be achieved asymptotically by m binary polar codes such that the channel input of the j-th user is given by $x_1^n[j] = u_1^n[j] G_n \in \{0,1\}^n$ for generator matrix G_n of binary polar codes. It is suggested in [22] that the complexity of Gallager's scheme may be improved by this technique since this technique uses $m = \log q$ binary polar codes instead of a single q-ary polar code. However, since the analysis of MAC polar codes heavily depends on the property of binary input, it is currently unknown whether polar codes over $\mathrm{GF}(\tilde{q})$ can be used for MAC coding to improve the exponent of the error probability.

Appendix

A. Proof of Theorem 2

Denote a member of $\tilde{\mathcal{Y}}^n = \{0,1\}^n \times \mathcal{Y}^n$ by $\tilde{y}_1^n = (\tilde{z}_1^n, y_1^n)$. Then we have

$$\begin{aligned}
&\tilde{W}_i^{(n)}(\tilde{u}_1^{i-1}, \tilde{y}_1^n | \tilde{u}_i) \\
&= P_{\tilde{U}_1^{i-1}, \tilde{Y}_1^n | \tilde{U}_i}(\tilde{u}_1^{i-1}, \tilde{y}_1^n | \tilde{u}_i) \\
&= \sum_{x_1^n} P_{X_1^n, Y_1^n, \tilde{X}_1^n, \tilde{U}_1^{i-1} | \tilde{U}_i}(x_1^n, y_1^n, \tilde{z}_1^n \oplus x_1^n, \tilde{u}_1^{i-1} | \tilde{u}_i) \\
&\overset{(a)}{=} \sum_{x_1^n} P_{X_1^n, Y_1^n}(x_1^n, y_1^n) P_{\tilde{X}_1^n, \tilde{U}_1^{i-1} | \tilde{U}_i}(\tilde{z}_1^n \oplus x_1^n, \tilde{u}_1^{i-1} | \tilde{u}_i) \\
&\overset{(b)}{=} 2 \sum_{x_1^n} P_{X_1^n, Y_1^n}(x_1^n, y_1^n) P_{\tilde{X}_1^n}(\tilde{z}_1^n \oplus x_1^n) \\
&\qquad \cdot \mathbb{1}\left[((\tilde{z}_1^n \oplus x_1^n) G_n)_1^i = \tilde{u}_1^i\right] \\
&\overset{(c)}{=} 2^{-n+1} \sum_{x_1^n} P_{X_1^n, Y_1^n}(x_1^n, y_1^n) \mathbb{1}\left[(x_1^n G_n)_1^i = \tilde{u}_1^i \oplus (\tilde{z}_1^n G_n)_1^i\right] \\
&\overset{(d)}{=} 2^{-n+1} P_{U_1^i, Y_1^n}(\tilde{u}_1^i \oplus (\tilde{z}_1^n G_n)_1^i, y_1^n), \qquad (34)
\end{aligned}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function, i.e. $\mathbb{1}[\mathrm{true}] = 1$ and $\mathbb{1}[\mathrm{false}] = 0$, and the equalities follow from
(a): $(\tilde{X}_1^n, \tilde{U}_1^n)$ is independent of (X_1^n, Y_1^n),
(b): $P_{\tilde{X}_1^n, \tilde{U}_1^{i-1} | \tilde{U}_i} = P_{\tilde{X}_1^n} P_{\tilde{U}_1^i | \tilde{X}_1^n} / P_{\tilde{U}_i}$ and $\tilde{U}_1^n = \tilde{X}_1^n G_n$,
(c): $\tilde{X}_1^n$ is uniformly distributed over $\{0,1\}^n$,
(d): $U_1^n = X_1^n G_n$.
We obtain (10) by letting $\tilde{z}_1^n = 0_1^n$.

Now we prove (11). From the definition of Z_{B} and (34) we have

$$\begin{aligned}
&Z_{\mathrm{B}}(\tilde{W}_i^{(n)}) \\
&= \sum_{\tilde{y}_1^n, \tilde{u}_1^{i-1}} \sqrt{P_{\tilde{U}_1^{i-1}, \tilde{Y}_1^n | \tilde{U}_i}(\tilde{u}_1^{i-1}, \tilde{y}_1^n | 0) P_{\tilde{U}_1^{i-1}, \tilde{Y}_1^n | \tilde{U}_i}(\tilde{u}_1^{i-1}, \tilde{y}_1^n | 1)} \\
&= \sum_{\tilde{z}_1^n, y_1^n, \tilde{u}_1^{i-1}} \sqrt{2^{-n+1} P_{U_1^i, Y_1^n}((\tilde{u}_1^{i-1}, 0) \oplus (\tilde{z}_1^n G_n)_1^i, y_1^n)} \\
&\qquad \cdot \sqrt{2^{-n+1} P_{U_1^i, Y_1^n}((\tilde{u}_1^{i-1}, 1) \oplus (\tilde{z}_1^n G_n)_1^i, y_1^n)} \\
&= 2^{-n+1} \sum_{\tilde{z}_1^n, y_1^n, \tilde{u}_1^{i-1}} \sqrt{P_{U_1^i, Y_1^n}((\tilde{u}_1^{i-1} \oplus (\tilde{z}_1^n G_n)_1^{i-1}, 0), y_1^n)} \\
&\qquad \cdot \sqrt{P_{U_1^i, Y_1^n}((\tilde{u}_1^{i-1} \oplus (\tilde{z}_1^n G_n)_1^{i-1}, 1), y_1^n)}. \qquad (35)
\end{aligned}$$

Let $z_1^n \equiv \tilde{z}_1^n G_n$ and $u_1^{i-1} \equiv \tilde{u}_1^{i-1} \oplus (\tilde{z}_1^n G_n)_1^{i-1} = \tilde{u}_1^{i-1} \oplus z_1^{i-1}$. Since $(\tilde{z}_1^n, \tilde{u}_1^{i-1}) \mapsto (z_1^n, u_1^{i-1})$ is a bijection on $\{0,1\}^n \times \{0,1\}^{i-1}$, it holds that

$$\begin{aligned}
&Z_{\mathrm{B}}(\tilde{W}_i^{(n)}) \\
&= 2^{-n+1} \sum_{z_1^n, y_1^n, u_1^{i-1}} \sqrt{P_{U_1^i, Y_1^n}((u_1^{i-1}, 0), y_1^n)} \\
&\qquad \cdot \sqrt{P_{U_1^i, Y_1^n}((u_1^{i-1}, 1), y_1^n)} \\
&= 2 \sum_{y_1^n, u_1^{i-1}} \sqrt{P_{U_1^i, Y_1^n}((u_1^{i-1}, 0), y_1^n) \cdot P_{U_1^i, Y_1^n}((u_1^{i-1}, 1), y_1^n)} \\
&= Z(U_i | U_1^{i-1}, Y_1^n) \qquad (36)
\end{aligned}$$

and the proof is completed. ■

B. Proof of Theorem 1

First we have

$$\begin{aligned} I(\tilde{W}) &= I(\tilde{X}; \tilde{X} \oplus X, Y) \\ &= H(\tilde{X} \oplus X, Y) - H(\tilde{X} \oplus X, Y|\tilde{X}) \\ &\overset{\text{(a)}}{=} 1 + H(Y) - H(X, Y|\tilde{X}) \\ &\overset{\text{(b)}}{=} 1 - H(X|Y)\,, \end{aligned} \tag{37}$$

where (a) holds since $\tilde{X}$ is uniformly distributed and independent of Y, and (b) follows from the independence between (X, Y) and $\tilde{X}$. Combining (37) with Theorem 2 and Proposition 2 we have

$$\lim_{n\to\infty} \frac{1}{n} \left| \left\{ i : Z(U_i|U_1^{i-1}, Y_1^n) \le 2^{-n^\beta} \right\} \right| = 1 - H(X|Y)\,, \tag{38}$$

$$\lim_{n\to\infty} \frac{1}{n} \left| \left\{ i : Z(U_i|U_1^{i-1}, Y_1^n) \ge 1 - 2^{-n^\beta} \right\} \right| = H(X|Y)\,. \tag{39}$$

Next consider the case that Y is a random variable which takes a fixed value with probability 1. For this case $Z(U_i|U_1^{i-1}, Y_1^n) = Z(U_i|U_1^{i-1})$ and $H(X|Y) = H(X)$. Thus, (38) and (39) become

$$\lim_{n\to\infty} \frac{1}{n} \left| \left\{ i : Z(U_i|U_1^{i-1}) \le 2^{-n^\beta} \right\} \right| = 1 - H(X)\,, \tag{40}$$

$$\lim_{n\to\infty} \frac{1}{n} \left| \left\{ i : Z(U_i|U_1^{i-1}) \ge 1 - 2^{-n^\beta} \right\} \right| = H(X)\,. \tag{41}$$

Let A–D be the sets of indices defined by

$$A \equiv \{ i : Z(U_i|U_1^{i-1}, Y_1^n) \le 2^{-n^\beta} \}\,, \tag{42}$$

$$B \equiv \{ i : Z(U_i|U_1^{i-1}, Y_1^n) \ge 1 - 2^{-n^\beta} \}\,, \tag{43}$$

$$C \equiv \{ i : Z(U_i|U_1^{i-1}) \le 2^{-n^\beta} \}\,, \tag{44}$$

$$D \equiv \{ i : Z(U_i|U_1^{i-1}) \ge 1 - 2^{-n^\beta} \}\,. \tag{45}$$

It is easy to see that $B \cap C$ is empty for sufficiently large n from Proposition 1 and $H(U_i|U_1^{i-1}, Y_1^n) \le H(U_i|U_1^{i-1})$. Furthermore, we also note from (38)–(41) that

$$\lim_{n\to\infty} \frac{|A| + |B|}{n} = \lim_{n\to\infty} \frac{|C| + |D|}{n} = 1\,. \tag{46}$$

Hence (5) and (6) hold because

$$\lim_{n\to\infty} \frac{|B \cup C|}{n} = \lim_{n\to\infty} \frac{|B| + |C|}{n} = 1 - I(X;Y) \tag{47}$$

and

$$\lim_{n\to\infty} \frac{|A \cap D|}{n} = 1 - \lim_{n\to\infty} \frac{|B \cup C|}{n} = I(X;Y)\,. \tag{48}$$

■

C. Proof of Theorem 3

Let $\mathcal{E}_i$ be the set of pairs of codeword $u_1^n = u_1^n(M_1^{|\mathcal{I}|}, \lambda_{\mathcal{I}^c})$ and received word y_1^n such that decoding error occurs at the ith bit. The block decoding error event is given by $\mathcal{E} \equiv \bigcup_{i\in\mathcal{I}} \mathcal{E}_i$. Under decoding given in (12) with an arbitrary tie-breaking rule, every $(u_1^n, y_1^n) \in \mathcal{E}_i$ satisfies

$$\begin{aligned} &P_{U_i|U_1^{i-1}, Y_1^n}(u_i|u_1^{i-1}, y_1^n) \\ &\quad \le P_{U_i|U_1^{i-1}, Y_1^n}(u_i \oplus 1|u_1^{i-1}, y_1^n)\,. \end{aligned} \tag{49}$$

Consider the block decoding error probability $P_{\mathrm{e}}(\lambda_{\mathcal{I}^c})$ for map $\lambda_{\mathcal{I}^c}$. Since each codeword u_1^n appears with probability

$$2^{-|\mathcal{I}|} \mathbb{1}\left[\bigcap_{i\in\mathcal{I}^c} \{\lambda_i(u_1^{i-1}) = u_i\} \right]\,, \tag{50}$$

$P_{\mathrm{e}}(\lambda_{\mathcal{I}^c})$ is given by

$$\begin{aligned} P_{\mathrm{e}}(\lambda_{\mathcal{I}^c}) = \sum_{u_1^n, y_1^n} 2^{-|\mathcal{I}|} \mathbb{1}\left[\bigcap_{i\in\mathcal{I}^c} \{\lambda_i(u_1^{i-1}) = u_i\} \right] \\ \cdot P_{Y_1^n|U_1^n}(y_1^n|u_1^n)\, \mathbb{1}[(u_1^n, y_1^n) \in \mathcal{E}]\,. \end{aligned} \tag{51}$$

From (13), the expectation of the decoding error probability is obtained as

$$\begin{aligned} \mathrm{E}_{\Lambda_{\mathcal{I}^c}}[P_{\mathrm{e}}(\Lambda_{\mathcal{I}^c})] = \sum_{u_1^n, y_1^n} 2^{-|\mathcal{I}|} \left(\prod_{i\in\mathcal{I}^c} P_{U_i|U_1^{i-1}}(u_i|u_1^{i-1}) \right) \\ \cdot P_{Y_1^n|U_1^n}(y_1^n|u_1^n)\, \mathbb{1}[(u_1^n, y_1^n) \in \mathcal{E}]\,. \end{aligned} \tag{52}$$

Then, using probability distribution $Q_{U_1^n, Y_1^n}$ defined as

$$\begin{aligned} &Q_{U_1^n, Y_1^n}(u_1^n, y_1^n) \\ &\equiv P_{Y_1^n|U_1^n}(y_1^n|u_1^n) 2^{-|\mathcal{I}|} \prod_{i\in\mathcal{I}^c} P_{U_i|U_1^{i-1}}(u_i|u_1^{i-1})\,, \end{aligned} \tag{53}$$

we can represent (52) as $\mathrm{E}_{\Lambda_{\mathcal{I}^c}}[P_{\mathrm{e}}(\Lambda_{\mathcal{I}^c})] = Q_{U_1^n, Y_1^n}(\mathcal{E})$. Let $\|F - G\|$ be the variational distance defined by

$$\begin{aligned} \|F - G\| &\equiv \frac{1}{2} \sum_x |F(x) - G(x)| \\ &= \sum_{x: F(x) > G(x)} (F(x) - G(x)) \end{aligned} \tag{54}$$

for probability distributions F and G. The variational distance between $Q_{U_1^n, Y_1^n}$ and $P_{U_1^n, Y_1^n}$ satisfies the following lemma.

Lemma 1. *For any $\beta < 1/2$ satisfying* (15) *and $\beta' < \beta$,*

$$\|P_{U_1^n, Y_1^n} - Q_{U_1^n, Y_1^n}\| = \mathrm{O}(2^{-n^{\beta'}})\,. \tag{55}$$

Proof: We use an argument similar to [6, Lemma 3.5] based on the expression

$$\begin{aligned} B_1^n - A_1^n &= \sum_{i=1}^n A_1^{i-1} B_i^n - \sum_{i=1}^n A_1^i B_{i+1}^n \\ &= \sum_{i=1}^n (B_i - A_i) A_1^{i-1} B_{i+1}^n \end{aligned} \tag{56}$$

where A_j^k and B_j^k denote products $\prod_{i=j}^k A_i$ and $\prod_{i=j}^k B_i$, respectively.

For simplicity, we omit the symbols of random variables, e.g. $P(u_1^n, y_1^n)$ and $Q(u_i|u_1^{i-1}, y_1^n)$ for $P_{U_1^n, Y_1^n}(u_1^n, y_1^n)$ and $Q_{U_i|U_1^{i-1}, Y_1^n}(u_i|u_1^{i-1}, y_1^n)$ in the following. Now $\|P_{U_1^n, Y_1^n} - Q_{U_1^n, Y_1^n}\|$ is bounded as (57), where $D(\cdot\|\cdot)$ is the relative entropy, and equality (a) and inequalities (b)–(d) follow from

(a): (56) and $Q(y_1^n|u_1^n) = P(y_1^n|u_1^n)$,
(b): $Q(u_i|u_1^{i-1}) = P(u_i|u_1^{i-1})$ for $i \in \mathcal{I}^c$,
(c): $\|F - G\| \le \sqrt{(\ln 2) D(F\|G)/2}$ by Pinsker's inequality (see, e.g., [23, Lemma 11.6.1]),
(d): Jensen's inequality.

(See below for Eq. (57).)

Hence, it holds that

$$\begin{aligned}
&2\|P_{U_1^n,Y_1^n} - Q_{U_1^n,Y_1^n}\| \\
&\le \sum_{i\in\mathcal{I}} \sqrt{(2\ln 2) D(P_{U_i}\|Q_{U_i}|U_1^{i-1})} \\
&\overset{\text{(e)}}{=} \sum_{i\in\mathcal{I}} \sqrt{(2\ln 2)(1 - H(U_i|U_1^{i-1}))} \\
&\overset{\text{(f)}}{\le} \sum_{i\in\mathcal{I}} \sqrt{(2\log 2)(1 - (Z(U_i|U_1^{i-1}))^2)} \\
&\overset{\text{(g)}}{\le} n\sqrt{(4\log 2)\cdot 2^{-n^\beta}} \\
&\overset{\text{(h)}}{=} \mathrm{O}(2^{-n^{\beta'}}),
\end{aligned} \tag{58}$$

where the equalities and the inequalities follow from (e): $Q_{U_i|U_1^{i-1}} = \frac{1}{2}$ for $i \in \mathcal{I}$, (f): Proposition 1, (g): (15) and (h): $\beta' < \beta$. ■

Proof of Theorem 3: First we have

$$\begin{aligned}
\mathrm{E}_{\Lambda_{\mathcal{I}^c}}[P_{\mathrm{e}}(\Lambda_{\mathcal{I}^c})] &= Q_{U_1^n,Y_1^n}(\mathcal{E}) \\
&\le \|Q_{U_1^n,Y_1^n} - P_{U_1^n,Y_1^n}\| + P_{U_1^n,Y_1^n}(\mathcal{E}) \\
&\le \|Q_{U_1^n,Y_1^n} - P_{U_1^n,Y_1^n}\| + \sum_{i\in\mathcal{I}} P_{U_1^n,Y_1^n}(\mathcal{E}_i).
\end{aligned} \tag{59}$$

Each term in the summation can be bounded as

$$\begin{aligned}
&P_{U_1^n,Y_1^n}(\mathcal{E}_i) \\
&\le \sum_{u_1^i,y_1^n} P(u_1^{i-1},y_1^n) P(u_i|u_1^{i-1},y_1^n) \\
&\qquad \cdot \mathbb{1}\left[P(u_i|u_1^{i-1},y_1^n) \le P(u_i\oplus 1|u_1^{i-1},y_1^n)\right] \\
&\le \sum_{u_1^i,y_1^n} P(u_1^{i-1},y_1^n) P(u_i|u_1^{i-1},y_1^n) \sqrt{\frac{P(u_i\oplus 1|u_1^{i-1},y_1^n)}{P(u_i|u_1^{i-1},y_1^n)}} \\
&= Z(U_i|U_1^{i-1},Y_1^n) \\
&\le 2^{-n^\beta},
\end{aligned} \tag{60}$$

where the last inequality follows from (15). From (55), (59) and (60), we have $\mathrm{E}_{\Lambda_{\mathcal{I}^c}}[P_{\mathrm{e}}(\Lambda_{\mathcal{I}^c})] = \mathrm{O}(2^{-n^{\beta'}})$. ■

D. Proof of Theorem 4

For a source sequence y_1^n and the encoding rule (19), $u_1^n = u_1^n(y_1^n, \lambda_{\mathcal{I}^c})$ appears with probability

$$\left(\prod_{i\in\mathcal{I}} P_{U_i|U_1^{i-1},Y_1^n}(u_i|u_1^{i-1},y_1^n)\right) \mathbb{1}\left[\bigcap_{i\in\mathcal{I}^c}\{\lambda_i(u_1^{i-1}) = u_i\}\right]. \tag{61}$$

The average distortion for map $\Lambda_{\mathcal{I}^c} = \lambda_{\mathcal{I}^c}$ is expressed as

$$\begin{aligned}
&D_n(\lambda_{\mathcal{I}^c}) \\
&= \frac{1}{n}\sum_{u_1^n,y_1^n} P_{Y_1^n}(y_1^n)\left(\prod_{i\in\mathcal{I}} P_{U_i|U_1^{i-1},Y_1^n}(u_i|u_1^{i-1},y_1^n)\right) \\
&\qquad \cdot \mathbb{1}\left[\bigcap_{i\in\mathcal{I}^c}\{\lambda_i(u_1^{i-1}) = u_i\}\right] d^n(y_1^n, u_1^n G_n)
\end{aligned} \tag{62}$$

$$\begin{aligned}
2\|P_{U_1^n,Y_1^n} - Q_{U_1^n,Y_1^n}\| &= \sum_{u_1^n,y_1^n} |Q(u_1^n,y_1^n) - P(u_1^n,y_1^n)| \\
&\overset{\text{(a)}}{=} \sum_{u_1^n,y_1^n} \left|\sum_i (Q(u_i|u_1^{i-1}) - P(u_i|u_1^{i-1}))\left(\prod_{j=1}^{i-1} P(u_j|u_1^{j-1})\right)\left(\prod_{j=i+1}^{N} Q(u_j|u_1^{j-1})\right) Q(y_1^n|u_1^n)\right| \\
&\overset{\text{(b)}}{\le} \sum_{i\in\mathcal{I}}\sum_{u_1^n,y_1^n} \left|Q(u_i|u_1^{i-1}) - P(u_i|u_1^{i-1})\right|\left(\prod_{j=1}^{i-1} P(u_j|u_1^{j-1})\right)\left(\prod_{j=i+1}^{N} Q(u_j|u_1^{j-1})\right) Q(y_1^n|u_1^n) \\
&= \sum_{i\in\mathcal{I}}\sum_{u_1^{i-1}} 2P(u_1^{i-1}) \|Q_{U_i|U_1^{i-1}=u_1^{i-1}} - P_{U_i|U_1^{i-1}=u_1^{i-1}}\| \\
&\overset{\text{(c)}}{\le} \sum_{i\in\mathcal{I}}\sum_{u_1^{i-1}} P(u_1^{i-1}) \sqrt{(2\ln 2) D(P_{U_i|U_1^{i-1}=u_1^{i-1}}\|Q_{U_i|U_1^{i-1}=u_1^{i-1}})} \\
&\overset{\text{(d)}}{\le} \sum_{i\in\mathcal{I}} \sqrt{(2\ln 2)\sum_{u_1^{i-1}} P(u_1^{i-1}) D(P_{U_i|U_1^{i-1}=u_1^{i-1}}\|Q_{U_i|U_1^{i-1}=u_1^{i-1}})}.
\end{aligned} \tag{57}$$

and its expectation over $\Lambda_{\mathcal{I}^c}$ is

$$\begin{aligned}&\mathrm{E}_{\Lambda_{\mathcal{I}^c}}[D_n(\Lambda_{\mathcal{I}^c})]\\&= \frac{1}{n}\sum_{u_1^n, y_1^n} P_{Y_1^n}(y_1^n)\left(\prod_{i\in\mathcal{I}} P_{U_i|U_1^{i-1},Y_1^n}(u_i|u_1^{i-1}, y_1^n)\right)\\&\quad\cdot\left(\prod_{i\in\mathcal{I}^c} P_{U_i|U_1^{i-1}}(u_i|u_1^{i-1})\right) d^n(y_1^n, u_1^n G_n)\,. \qquad (63)\end{aligned}$$

Then, for probability distribution $Q_{U_1^n,Y_1^n}$ defined as

$$\begin{aligned}Q_{U_1^n,Y_1^n}(u_1^n, y_1^n) \equiv P_{Y_1^n}(y_1^n)\left(\prod_{i\in\mathcal{I}} P_{U_i|U_1^{i-1},Y_1^n}(u_i|u_1^{i-1}, y_1^n)\right)\\\cdot\left(\prod_{i\in\mathcal{I}^c} P_{U_i|U_1^{i-1}}(u_i|u_1^{i-1})\right), \qquad (64)\end{aligned}$$

(63) is represented as

$$\mathrm{E}_{\Lambda_{\mathcal{I}^c}}[D_n(\Lambda_{\mathcal{I}^c})] = \frac{1}{n}\mathrm{E}_{Q_{U_1^n,Y_1^n}}[d(Y_1^n, U_1^n G_n)]\,. \qquad (65)$$

Therefore we obtain

$$\begin{aligned}\mathrm{E}_{\Lambda_{\mathcal{I}^c}}[D_n(\Lambda_{\mathcal{I}^c})] \le \frac{1}{n}\mathrm{E}_{P_{U_1^n,Y_1^n}}[d^n(Y_1^n, G_n U_1^n)]\\+ \frac{\max_{y,x} d(y,x)}{n}\|P_{U_1^n,Y_1^n} - Q_{U_1^n,Y_1^n}\| \qquad (66)\end{aligned}$$

and the following lemma shows that the second term of the RHS of (66) is $\mathrm{O}(2^{-n^{\beta'}})$. ■

Lemma 2. *For any $\beta < 1/2$ satisfying* (18) *and $\beta' < \beta$,*

$$\|P_{U_1^n,Y_1^n} - Q_{U_1^n,Y_1^n}\| = \mathrm{O}(2^{-n^{\beta'}})\,. \qquad (67)$$

Proof: By the same argument and notation as the proof of Lemma 1, $\|P_{U_1^n,Y_1^n} - Q_{U_1^n,Y_1^n}\|$ is bounded as (68), where the equalities and the inequality follow from

(a): (56) and $Q(y_1^n) = P(y_1^n)$,
(b): $Q(u_i|u_1^{i-1}, y_1^n) = P(u_i|u_1^{i-1}, y_1^n)$ for $i \in \mathcal{I}$,
(c): $Q_{U_i|U_1^{i-1},Y_1^n} = P_{U_i|U_1^{i-1}}$ for $i \in \mathcal{I}^c$.

(See below for Eq. (68).)

Furthermore it holds for all $i \in \mathcal{I}^c$ that

$$\begin{aligned}&H(U_i|U_1^{i-1}) - H(U_i|U_1^{i-1}, Y_1^n)\\&\overset{(d)}{\le} Z(U_i|U_1^{i-1}) - (Z(U_i|U_1^{i-1}, Y_1^n))^2\\&\overset{(e)}{\le} \min\{Z(U_i|U_1^{i-1}),\, 1 - (Z(U_i|U_1^{i-1}, Y_1^n))^2\}\\&\overset{(f)}{\le} 2\cdot 2^{-n^\beta} \qquad (69)\end{aligned}$$

from (d): Proposition 1, (e): $Z(\cdot|\cdot) \in [0,1]$ and (f): (18). We obtain the lemma by combining (68) and (69). ■

Acknowledgment

The authors are grateful to Dr. Ryuhei Mori for helpful discussions. They also thank the associate editor and reviewers for their valuable comments.

References

[1] E. Arikan, "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] R. Mori and T. Tanaka, "Channel polarization on q-ary discrete memoryless channels by arbitrary kernels," in *Proceedings of IEEE International Symposium on Information Theory (ISIT10)*, 2010, pp. 894–898.

[3] E. Şaşoğlu, E. Telatar, and E. Arikan, "Polarization for arbitrary discrete memoryless channels," in *Proceedings of IEEE Information Theory Workshop (ITW2009)*, 2009, pp. 144–148.

[4] H. Cronie and S. Korada, "Lossless source coding with polar codes," in *Proceedings of IEEE International Symposium on Information Theory (ISIT10)*, 2010, pp. 904–908.

$$\begin{aligned}&2\|P_{U_1^n,Y_1^n} - Q_{U_1^n,Y_1^n}\|\\&\overset{(a)}{=} \sum_{u_1^n,y_1^n}\left|\sum_i (Q(u_i|u_1^{i-1}, y_1^n) - P(u_i|u_1^{i-1}, y_1^n))P(y_1^n)\left(\prod_{j=1}^{i-1} P(u_j|u_1^{j-1}, y_1^n)\right)\left(\prod_{j=i+1}^{N} Q(u_j|u_1^{j-1}, y_1^n)\right)\right|\\&\overset{(b)}{\le} \sum_{i\in\mathcal{I}^c}\sum_{u_1^i,y_1^n}\left|Q(u_i|u_1^{i-1}, y_1^n) - P(u_i|u_1^{i-1}, y_1^n)\right| P(y_1^n)\left(\prod_{j=1}^{i-1} P(u_j|u_1^{j-1}, y_1^n)\right)\\&= \sum_{i\in\mathcal{I}^c}\sum_{u_1^{i-1},y_1^n} 2P(u_1^{i-1}, y_1^n)\|Q_{U_i|Y_1^n=y_1^n, U_1^{i-1}=u_1^{i-1}} - P_{U_i|Y_1^n=y_1^n, U_1^{i-1}=u_1^{i-1}}\|\\&\le \sum_{i\in\mathcal{I}^c}\sum_{u_1^{i-1},y_1^n} P(u_1^{i-1}, y_1^n)\sqrt{(2\ln 2)D(P_{U_i|Y_1^n=y_1^n, U_1^{i-1}=u_1^{i-1}}\|Q_{U_i|Y_1^n=y_1^n, U_1^{i-1}=u_1^{i-1}})}\\&\le \sum_{i\in\mathcal{I}^c}\sqrt{(2\ln 2)\sum_{u_1^{i-1},y_1^n} P(u_1^{i-1}, y_1^n)D(P_{U_i|Y_1^n=y_1^n, U_1^{i-1}=u_1^{i-1}}\|Q_{U_i|Y_1^n=y_1^n, U_1^{i-1}=u_1^{i-1}})}\\&= \sum_{i\in\mathcal{I}^c}\sqrt{(2\ln 2)D(P_{U_i}\|Q_{U_i}|U_1^{i-1}, Y_1^n)}\\&\overset{(c)}{=} \sum_{i\in\mathcal{I}^c}\sqrt{(2\ln 2)(H(U_i|U_1^{i-1}) - H(U_i|U_1^{i-1}, Y_1^n))}\,. \qquad (68)\end{aligned}$$

[5] S. Korada and R. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1751–1768, 2010.
[6] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, Lausanne, 2009. [Online]. Available: http://library.epfl.ch/theses/?nr=4461
[7] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[8] E. Arikan, "Source polarization," in *Proceedings of IEEE International Symposium on Information Theory (ISIT10)*, 2010, pp. 899–903.
[9] D. Sutter, J. M. Renes, F. Dupuis, and R. Renner, "Achieving the capacity of any DMC using only polar codes," in *Proceedings of IEEE Information Theory Workshop (ITW2012)*, Lausanne, Switzerland, Sep. 2012, pp. 114–118. [Online]. Available: http://arxiv.org/abs/1205.3756v2
[10] I. Tal and A. Vardy, "How to construct polar codes," *submitted to IEEE Trans. Inform. Theory*, 2011. [Online]. Available: http://arxiv.org/abs/arXiv:1105.6164v2
[11] E. Arikan and E. Telatar, "On the rate of channel polarization," in *Proceedings of IEEE International Symposium on Information Theory (ISIT09)*, 2009, pp. 1493–1495.
[12] G. P. McCormick, "Second order conditions for constrained minima," *SIAM Journal on Applied Mathematics*, vol. 15, pp. 641–652, 1967.
[13] E. Abbe and A. Barron, "Polar coding schemes for the AWGN channel," in *Proceedings of IEEE International Symposium on Information Theory (ISIT11)*, 2011, pp. 194–198.
[14] Y. Wu and S. Verdú, "The impact of constellation cardinality on Gaussian channel capacity," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010, pp. 620–628.
[15] J. G. Smith, "The information capacity of amplitude- and variance-constrained scalar Gaussian channels," *Information and Control*, pp. 203–219, 1971.
[16] T. H. Chan, S. Hranilovic, and F. R. Kschischang, "Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs," *IEEE Trans. Inform. Theory*, vol. 51, no. 6, pp. 2073–2088, 2005.
[17] S. Hassani and R. Urbanke, "On the scaling of polar codes: I. the behavior of polarized channels," in *Proceedings of IEEE International Symposium on Information Theory (ISIT10)*, June, pp. 874–878.
[18] T. Tanaka and R. Mori, "Refined rate of channel polarization," in *Proceedings of IEEE International Symposium on Information Theory (ISIT10)*, June, pp. 889–893.
[19] S. Hassani, R. Mori, T. Tanaka, and R. Urbanke, "Rate-dependent analysis of the asymptotic behavior of channel polarization," *IEEE Trans. Inform. Theory*, vol. 59, no. 4, pp. 2267–2276, 2013.
[20] R. Mori and T. Tanaka, "Source and channel polarization over finite fields and Reed-Solomon matrix," submitted to *IEEE Trans. Inform. Theory*. [Online]. Available: http://arxiv.org/abs/1110.0194
[21] S. 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, 2010.
[22] E. Abbe and I. Telatar, "Polar codes for the m-user multiple access channel," *IEEE Trans. Inform. Theory*, vol. 58, no. 8, pp. 5437–5448, 2012.
[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, July 2006.

Junya Honda was born in Kanagawa, Japan, in 1985. He received the B.E., M.E. and Ph.D. degrees from the University of Tokyo, Tokyo, Japan in 2008, 2010 and 2013, respectively. Since 2013, he has been a Research Associate at the University of Tokyo and is currently with the department of Complexity Science and Engineering at the university. His research interests are in Shannon theory, coding theory and statistical learning theory.

Hirosuke Yamamoto (S'77–M'80–SM'03–F'11) was born in Wakayama, Japan, in 1952. He received the B.E. degree from Shizuoka University, Shizuoka, Japan, in 1975 and the M.E. and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1977 and 1980, respectively, all in electrical engineering. In 1980, he joined Tokushima University. He was an Associate Professor at Tokushima University from 1983 to 1987, the University of Electro-Communications from 1987 to 1993, and the University of Tokyo from 1993 to 1999. Since 1999, he has been a Professor at the University of Tokyo and is currently with the department of Complexity Science and Engineering at the university. In 1989–1990, he was a Visiting Scholar at the Information Systems Laboratory, Stanford University, Stanford, CA. His research interests are in Shannon theory, data compression algorithms, and cryptology.

Dr. Yamamoto served as the Chair of IEEE Information Theory Society Japan Chapter in 2002–2003, the TPC Co-Chair of the ISITA2004, the TPC Chair of the ISITA2008, the president of the SITA (Society of Information Theory and its Applications) in 2008–2009, the president of the ESS (Engineering Sciences Society) of IEICE in 2012–2013, an Associate Editor for Shannon Theory, the IEEE TRANSACTIONS ON INFORMATION THEORY in 2007–2010, Editor-in-Chief for the IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences in 2009–2011. He is a Fellow of the IEICE.