Properties of a certain stochastic dynamical system, channel polarization, and polar codes

Toshiyuki Tanaka
Graduate School of Informatics, Kyoto University, Yoshida Hon-machi, Sakyo-ku, Kyoto-shi, Kyoto 606-8501, Japan
E-mail: tt01.kyoto-u.ac.jp

Abstract. A new family of codes, called polar codes, has recently been proposed by Arıkan. Polar codes are of theoretical importance because they are provably capacity achieving with low-complexity encoding and decoding. We first discuss basic properties of a certain stochastic dynamical system, on the basis of which properties of channel polarization and polar codes are reviewed, with emphasis on our recent results.

1. Introduction

The concept “More is different” is relevant not only in statistical physics but also in various other research fields, such as information theory, where one can find several examples in which properties of a system drastically change as the system size becomes large. As an example, channel coding theorem states that there exist error-correcting codes such that their error probabilities are arbitrarily small and yet their coding rates are arbitrarily close to the channel capacity. Although channel coding theorem is one of the most important theorems in information theory, it is also widely understood that channel coding theorem is a theorem of existence, so that one requires different arguments as to how one can construct practical codes with high performance. The problem of constructing codes which have practically low computational complexity in encoding and decoding, as well as good performance in view of channel coding theorem, remains a problem of particular importance in information theory and coding theory.

Proposal of turbo codes in 1990s, as well as subsequent rediscovery of low-density parity-check (LDPC) codes, has ignited extensive research activities on the combination of codes with sparse graphical representation and probability-based iterative decoding algorithms. As a consequence, by now we have practical means of how to construct codes with low computational complexity and high performance. However, rigorous results justifying the high performance of LDPC codes are mostly limited to the case with erasure channels, and we have not arrived at a thorough understanding of LDPC codes for other channels.

Arıkan [1] recently proposed polar codes as a class of error-correcting codes which is completely different from LDPC codes. Polar codes are theoretically very interesting because it saturates symmetric channel capacity (maximum mutual information when input distribution is assumed uniform, which is equal to the channel capacity for a symmetric channel) in the large codelength limit, whereas computational complexity of encoding and decoding is polynomial in the codelength. In this paper, we review recent developments regarding polar codes, as well as
channel polarization which forms mathematical foundation of polar codes, putting emphasis on our own results.

2. Properties of a stochastic dynamical system

In this section we define a stochastic dynamical system and argue its basic properties. The argument in this section will be utilized in understanding fundamentals of channel polarization and polar codes.

Let $f_0$ and $f_1$ be functions on the interval $[0, 1]$ defined as

$$f_0(x) = 1 - (1 - x)^2,$$
$$f_1(x) = x^2.$$  \hfill (1)
stochastic dynamical system as follows.

\[ p_0 = p \in (0, 1) \]
\[ p_n = f_{B_n}(p_{n-1}) = \begin{cases} f_0(p_{n-1}), & \text{if } B_n = 0 \\ f_1(p_{n-1}), & \text{if } B_n = 1 \end{cases} \tag{2} \]

In other words, the stochastic dynamical system is defined by the stochastic updating rule in which one throws a fair coin at every discrete time, and updates the state, denoted by \( p_n \) indexed by discrete time \( n \), with either of the functions \( f_0 \) and \( f_1 \) according to the outcome of the coin flip. There are \( 2^n \) possible values of \( p_n \) with equal probability, and a realization of \( p_n \) depends on a realization of \( \{B_1, B_2, \ldots, B_n\} \), so that \( \{p_0, p_1, \ldots, \} \) is a stochastic process. Figure 1 shows a sample process that can be obtained from this dynamical process. Figure 2 shows all possible trajectories from the initial value \( p = 0.3 \).

For the stochastic dynamical system (2), it is not difficult to show the following proposition \([1]\) to hold.

**Proposition 1** The limit \( \lim_{n \to \infty} p_n = p_\infty \) exists, and

\[ P(p_\infty) = \begin{cases} 1 - p & (p_\infty = 0) \\ p & (p_\infty = 1) \end{cases} \tag{3} \]

holds.

We show an outline of the proof, essentially due to Arikan and Telatar \([2]\), in the following. First, from the identity \( [f_0(x) + f_1(x)]/2 = x \), one can prove \( \mathbb{E}(p_n|p_{n-1}) = p_{n-1} \). It means that the stochastic process \( \{p_0, p_1, \ldots, \} \) defined by the stochastic dynamical system (2) is a martingale. It is obvious that this martingale is bounded, so that one can apply the martingale convergence theorem to prove that \( p_n \) converges in the limit \( n \to \infty \) almost surely to a random variable \( p_\infty \). It is also obvious that \( \mathbb{E}(p_\infty) = p \) holds.

Next, in order to study properties of the limiting random variable \( p_\infty \), let us define a stochastic process \( \{q_0, q_1, \ldots\} \) by

\[ q_n := p_n(1 - p_n). \tag{4} \]

For this stochastic process, one has

\[ q_n = q_{n-1} \times \begin{cases} (2 - p_{n-1})(1 - p_{n-1}) & \text{if } B_n = 0, \\ p_{n-1}(1 + p_{n-1}) & \text{if } B_n = 1. \end{cases} \tag{5} \]

On the other hand, from Schwartz inequality, one has, for \( \forall x \in [0, 1], \)

\[ \left[ \sqrt{(2 - x)(1 - x)} + \sqrt{x(1 + x)} \right]^2 \leq [(2 - x) + (1 + x)] \cdot [(1 - x) + x] = 3, \tag{6} \]

which proves

\[ \mathbb{E}(q_n^{1/2} | q_{n-1}) \leq \sqrt{\frac{3}{4} q_{n-1}^{1/2}}. \tag{7} \]

Recursive application of the above inequality gives

\[ \mathbb{E}(q_n^{1/2}) \leq \frac{1}{2} \left( \frac{3}{4} \right)^{n/2}. \tag{8} \]

Since \( q_n \) is non-negative, (8) implies that \( q_n \) takes values close to 0 with probability close to 1. More precisely, application of Markov inequality yields, for any \( \rho > 0 \),

\[ P \left( q_n \geq \rho n \right) \leq \left( \frac{3}{4} \rho \right)^{n/2} \tag{9} \]
holds. Let $\rho \in (3/4, 1)$, then the probability of the event $q_n < \rho^n/4$ approaches 1 exponentially as $n$ tends to large. That $q_n$ approaches 0 corresponds to $p_n$ approaching 0 or 1. More explicitly, one has

$$P(p_n < \rho^n) + P(p_n > 1 - \rho^n) \geq P\left(q_n < \frac{\rho^n}{4}\right) \geq 1 - \left(\frac{3}{4\rho}\right)^{n/2},$$

proving that $p_\infty$ takes 0 or 1 with probability 1. This observation, along with the fact $E(p_\infty) = p$, proves Proposition 1.

Figure 3 shows cumulative distributions of $p_n$ with the initial condition $p = 0.3$, for the cases with $n = 5, 10, 15, 20$. It can be observed that the distributions approach the limiting distribution as stated in Proposition 1.

3. Channel polarization and polar codes

3.1. Successive cancellation decoding

We first construct a code with codeword length 2 and code rate 1, which forms a basis for polar codes. The code is a linear code on $\text{GF}(2)$ (Galois field of order 2), whose generator matrix is

$$F = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \tag{11}$$

In other words, letting $(u_0, u_1) \in \{0, 1\}^2$ be a message (for later convenience, we adopt the convention to start the index of a vector from 0 rather than from 1), encoding is represented as an operation of $F$ from right

$$(u_0, u_1)F = (u_0, u_1) \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = (u_0 + u_1, u_1), \tag{12}$$

producing a codeword $(x_0, x_1) = (u_0 + u_1, u_1)$. Here, $u_0 + u_1$ means the addition in $\text{GF}(2)$.

Assume that every symbol of the codeword is to be transmitted over a binary erasure channel (BEC) with erasure probability $p$. Let $(y_0, y_1)$ be the received word, then $y_0$ is equal to $x_0$ with probability $(1 - p)$, and an erasure symbol with probability $p$. Similarly, $y_1$ is equal to $x_1$ with probability $(1 - p)$, and an erasure symbol with probability $p$. In addition, $y_0$ and $y_1$ are independent given $(x_0, x_1)$.
The problem of decoding here is to estimate \((u_0, u_1)\) on the basis of the received word \((y_0, y_1)\). Although there would be several ways to do the decoding, we consider in the following the decoding in which \(u_0\) and \(u_1\) are estimated individually and sequentially. We first consider the estimation of \(u_0\). The estimation of \(u_0\) is not possible if either \(y_0\) or \(y_1\) is erased: If \(y_0\) is unknown, information about \(u_0\) is no longer available. Even if \(y_0 = x_0 = u_0 + u_1\) is available, estimation of \(u_0\) is not possible if \(y_1 = u_1\) is unknown. The probability \(p'\) of failure in estimating \(u_0\) is, therefore, equal to the probability that either \(y_0\) or \(y_1\) (or both) is erased, and is given by \(f_0(p) = 1 - (1 - p)^2\). Figure 4 (a) depicts a factor-graph representation of the problem of estimating \(u_0\).

We next consider the problem of estimating \(u_1\), assuming that we have the correct estimation result for \(u_0\). One can estimate \(u_1\) correctly if \(y_1 = u_1\) is not erased. Or even though \(y_1\) is erased, one can still estimate \(u_1\) correctly if \(y_0 = u_0 + u_1\) is not erased, by making use of the knowledge of \(u_0\) which is assumed to be correctly estimated. Thus, the probability \(p''\) that one cannot correctly estimate \(u_1\) on the assumption that \(u_0\) has been correctly estimated is equal to the probability that \(y_0\) and \(y_1\) are both erased, and therefore given by \(f_1(p) = p^2\). The problem of estimating \(u_1\) is represented by a factor graph as in figure 4 (b). The approach of decoding discussed so far, in which the components \(u_0\) and \(u_1\) are estimated one-by-one sequentially is called successive cancellation decoding [1].

On the basis of the above-defined code, we next construct a code with codelength 4 and code rate 1. It is a linear code over GF(2), as before, and is constructed so that the problem of decoding a codeword \((u_0, u_1, u_2, u_3)\) from a received word \((y_0, y_1, y_2, y_3)\) has a factor-graph representation as shown in figure 5.

Let us consider successive cancellation decoding of this code applied to a binary erasure channel with erasure probability \(p\). Remembering the above argument on successive cancellation decoding of the code with codelength 2, in order to correctly estimate \(u_0\), one needs correct estimation results for the intermediate variables \(v_0\) and \(v_2\) in figure 5. Since the probabilities
that $v_0$ and $v_2$ are not correctly estimated are both equal to $p' = f_0(p)$, the probability that $u_0$ is not correctly estimated is equal to $1 - (1-p')^2 = f_0(f_0(p))$. Next, we consider the probability that one cannot correctly estimate $u_1$ under the assumption that one knows the correct estimation result for $u_0$. The probability is equal to the probability that neither $v_0$ nor $v_2$ is correctly estimated, and is thus given by $(p')^2 = f_1(f_0(p))$. Similarly, the probability that $u_2$ cannot be correctly estimated provided that the correct estimates of $u_0$ and $u_1$ are known is equal to $f_0(f_1(p))$, and the probability that $u_3$ cannot be correctly estimated provided that the correct estimates of $u_0$, $u_1$, and $u_2$ are known is given by $f_1(f_1(p))$.

In a similar way, we can recursively construct codes whose codelengths are integer powers of 2 (see figure 6): Let $G_2 = F$, and construct the generator matrices recursively as

$$G_N = (I_{N/2} \otimes F) R_N (I_2 \otimes G_{N/2}).$$

The matrix $R_N$ is called a reverse-shuffle operator which, when applied to a vector from right, induces a permutation among elements of the vector that corresponds to the inverse of the so-called perfect shuffle of a deck of playing cards (see figure 7). For example,

$$R_8 = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{pmatrix}.$$
For the successive cancellation decoding of the code of codelength \( N = 2^n \) and a binary erasure channel with erasure probability \( p \), the probability of not correctly estimating \( u_0 \) is given by

\[
\frac{f_0 \circ f_0 \circ \cdots \circ f_0(p)}{n \text{ operations}}
\tag{15}
\]

and, assuming that we know the correct estimation results for \( u_0, \ldots, u_{i-1} \), the probability that one fails to estimate \( u_i \) is given by

\[
\frac{f_{b_1} \circ f_{b_2} \circ \cdots \circ f_{b_n}(p)}{n \text{ operations}}
\tag{16}
\]

where \( b_n b_{n-1} \ldots b_2 b_1 \) denotes binary expansion of \( i \).

### 3.2. Polar codes

The codes defined in the previous subsection has code rate 1, so that they cannot be used as error-correcting codes as they are. One can construct codes with code rate less than 1 from the above rate-1 codes, by simply clamping some components of \( u = (u_0, u_1, \ldots, u_{N-1}) \) to 0, which is called shortening of codes. Let \( \mathcal{I} \) denote the set of indices of unclamped components of \( u \). A component \( u_i \) with \( i \notin \mathcal{I} \) is called a frozen bit. A polar code is constructed by appropriately choosing \( \mathcal{I} \) to make error probability of successive cancellation decoding small.

One could consider a strategy that defines \( \mathcal{I} \) as a set of indices of the rows of \( G_N \) containing the largest numbers of 1s. The resulting codes are equivalent to the Reed-Muller codes. Polar codes are constructed on the basis of a different strategy to define \( \mathcal{I} \), as described below.

In performing successive cancellation decoding, one does not have to decode the frozen bits \( \{u_i; i \notin \mathcal{I}\} \) since one knows that these values have been clamped to 0. Let \( \hat{u} \) be the decoding result of successive cancellation decoding. Let \( \mathcal{E}_N \) be the event that successive cancellation decoding fails in a polar code with codelength \( N \), and let \( B_{i, N} \) for \( i \in \mathcal{I} \) be the event that successive cancellation decoding fails for the first time at the \( i \)th component. Then, \( \{B_{i, N}; i \in \mathcal{I}\} \) are mutually exclusive, and

\[
\mathcal{E}_N = \bigcup_{i \in \mathcal{I}} B_{i, N}
\tag{17}
\]

holds. Let \( A_{i, N} \) be the event that successive cancellation decoding fails for \( u_i \) provided that the decoder knows the correct decoding results for \( u_0^{i-1} \), where and hereafter, we adopt the notational convention of expressing a subvector as \( u_i^j = (u_i, u_{i+1}, \ldots, u_{j-1}, u_j) \), etc. Noting that \( B_{i, N} \subset A_{i, N} \), one obtains an upper bound of the error probability of successive cancellation decoding as

\[
P(\mathcal{E}_N) = \sum_{i \in \mathcal{I}} P(B_{i, N}) \leq \sum_{i \in \mathcal{I}} P(A_{i, N}).
\tag{18}
\]

Arikan [1] proposed constructing a polar code by choosing \( \mathcal{I} \) to minimize the upper bound (18).

Since the resulting code is a linear code, computational complexity of encoding is at most \( O(N^2) \): In fact, one can exploit the recursive structure of the generator matrix to reduce the complexity of encoding to \( O(N \log N) \). On the other hand, computational complexity of successive cancellation decoding is \( O(N) \) per component of \( u \), but the total complexity can be reduced to \( O(N \log N) \) since intermediate results for decoding of one component can be utilized in decoding of other components. Computational complexity of polar codes is therefore \( O(N \log N) \) for encoding as well as for successive cancellation decoding.
3.3. Channel polarization

Knowledge of the probabilities \( \{P(A_{i,N})\} \) is essential in order to construct polar codes on the basis of the method given in the previous subsection. It is also important when one wants to know performance of polar codes.

As already seen in section 3.1, the problem of evaluating the values of the probabilities \( \{P(A_{i,N})\} \) has been completely solved: Let \( b_n b_{n-1} \ldots b_2 b_1 \) be the binary expansion of \( i \), then

\[
P(A_{i,N}) = f_{b_1} \circ f_{b_2} \circ \cdots \circ f_{b_n}(p)
\]

(19)

holds. Moreover, on the basis of the results reviewed in section 2, one immediately has, as \( N = 2^n \) tends to be large, that \( \{P(A_{i,N})\} \) will take values close to 0 or 1 with high probability, and that fractions of indices \( i \) for which \( P(A_{i,N}) \) takes a value close to 0 and 1 are \((1-p)\) and \( p \), respectively. Arıkan called such a phenomenon channel polarization. From the above-mentioned facts, the number of indices \( i \) such that \( P(A_{i,N}) \) take values close to 0 is approximately \( N(1-p) \) as \( N \) gets large, one can collect those indices to define \( I \) to construct a polar code with code rate less than \((1-p)\) and a vanishing error probability. In order to demonstrate that the above scenario is justified, one has to prove that the error probability \( P(\mathcal{E}_N) \) is \( o(1) \) by showing that the fraction of indices \( i \) for which \( P(A_{i,N}) = o(1/N) \) hold can be made arbitrarily close to \((1-p)\).

Intuitively, if \( p_n \) takes a value close to 0, then it is very unlikely that \( p_n \) takes a large value for any \( n \gg m \). In order to see this observation in a more concrete setting, we define a stochastic process \( \{\tilde{p}_0, \tilde{p}_1, \ldots\} \) as

\[
\tilde{p}_n = p_n \quad (n \leq m)
\]

\[
\tilde{p}_n = \begin{cases} 
2\tilde{p}_{n-1}, & \text{if } B_n = 0, \\
\tilde{p}_{n-1}^2, & \text{if } B_n = 1.
\end{cases} \quad (n > m)
\]

(20)

(21)

Obviously \( \tilde{p}_n \geq p_n \) holds because of the relation \( 2x \geq f_0(x) \). Taking logarithm of both sides of (21) yields

\[
\log_2 \tilde{p}_n = \begin{cases} 
\log_2 \tilde{p}_{n-1} + 1, & \text{if } B_n = 0, \\
2\log_2 \tilde{p}_{n-1}, & \text{if } B_n = 1.
\end{cases} \quad (n > m)
\]

(22)

One observes that the addition by 1 when \( B_n = 0 \) has a relatively negligible effect if \( \tilde{p}_m \) is sufficiently close to 0, because \( \log_2 \tilde{p}_{n-1} \) should take large negative values. Therefore, from the law of large numbers, one typically has

\[
\tilde{p}_n \approx p_m^{2(n-m)/2}.
\]

(23)

More precisely, the probability that

\[
p_n = o \left( 2^{-2^{3\beta n}} \right) = o \left( 2^{-N^{3\beta}} \right)
\]

(24)

holds for any \( \beta < 1/2 \) tends to \((1-p)\) as \( n \to \infty \), which implies that block error probability \( P(\mathcal{E}_N) \) satisfies

\[
P(\mathcal{E}_N) = o \left( 2^{-2^{3n}} \right)
\]

(25)

whenever the code rate \( R \) is less than \((1-p)\). This result is due to Arıkan and Telatar [2]. We have further refined the above result by applying the central limit theorem, to obtain the following result depending on the code rate \( R \) [3, 4]:

\[
P(\mathcal{E}_N) = o \left( 2^{-2^{(n+t+\sqrt{t}^2)/2}} \right) \quad \text{for any } t \text{ satisfying } Q(t) > \frac{R}{1-p}
\]

(26)
Here $Q(t) = \int_{-\infty}^{\infty} e^{-u^2/2} du / \sqrt{2\pi}$ is the conventional Q-function.

It should be noted that $(1-p)$ is nothing but the channel capacity of a binary erasure channel with erasure probability $p$. The argument so far has therefore proved that polar codes can achieve the channel capacity of binary erasure channels asymptotically in the large codelength limit, and that the error probability of successive cancellation decoding decays like $o\left(2^{-2^{n+Q^{-1}(R/(1-p))}n^2/2}\right)$ provided that the code rate $R$ is less than the channel capacity $(1-p)$.

### 3.4. Polarization for general binary-input memoryless channel

We have so far discussed channel polarization and polar codes assuming binary erasure channels. Even if one assumes a general binary-input memoryless channel, one can construct polar codes which asymptotically saturates the symmetric capacity on the basis of channel polarization.

A binary-input memoryless channel is characterized by the conditional probability $W(y|x)$ of output $y$ given input $x$, where $x \in \{0, 1\}$. We first discuss the combination of the code with the generator matrix $F$ defined in (11), which has codelength 2 and code rate 1, and successive cancellation decoding, as discussed in section 3.1. Assuming that $u_0$ and $u_1$ take 0 or 1 independently with probability 1/2, one can regard that the combination of the code and successive cancellation decoding decoding generates, from the channel $W$, a pair of two channels defined by

$$W_2^0(y_0, y_1|u_0) = \sum_{u_1=0,1} \frac{1}{2} W(y_0|u_0 + u_1)W(y_1|u_1),$$

$$W_2^1(y_0, y_1, u_0|u_1) = \frac{1}{2} W'(y_0|u_0 + u_1)W(y_1|u_1).$$

(27)

The factor 1/2 on the right-hand sides of the above equations corresponds to the prior probabilities of $u_1$ and $u_0$, respectively. In successive cancellation decoding, one first obtains an estimation result $\hat{u}_0$ from $(y_0, y_1)$ via maximum-likelihood decoding of the channel $W_2^0(y_0, y_1|u_0)$, as

$$\hat{u}_0 = \text{arg max}_{u_0=0,1} W_2^0(y_0, y_1|u_0).$$

(28)

Assuming that the estimation result be correct, one next obtains an estimate $\hat{u}_1$ from $(y_0, y_1)$ and $u_0$ via maximum-likelihood decoding of the channel $W_2^1(y_0, y_1, u_0|u_1)$, as

$$\hat{u}_1 = \text{arg max}_{u_1=0,1} W_2^1(y_0, y_1, u_0|u_1).$$

(29)

One can extend the above transformation of channels in accordance with the recursive construction of codes of codelength $N = 2^n$ described in section 3.1, and define the recursive process of generating pairs of channels $(W_{N}^{2i}, W_{N}^{2i+1})$ from $W_{N/2}^i$ starting with the given channel $W(y|x) = W_1^0(y|x)$, as

$$W_{N}^{2i}(y_0^{N-1}, u_0^{2i-1}|u_{2i}) = \sum_{u_{2i+1}=0,1} \frac{1}{2} W_{N/2}^i(y_0^{N/2-1}, u_0^{2i-1} + u_0^{2i-1}|u_{2i} + u_{2i+1})W_{N/2}^i(y_{N/2}^{N-1}, u_0^{2i-1}|u_{2i+1}),$$

(30)

$$W_{N}^{2i+1}(y_0^{N-1}, u_0^{2i}|u_{2i+1}) = \frac{1}{2} W_{N/2}^i(y_0^{N/2-1}, u_0^{2i-1} + u_0^{2i-1}|u_{2i} + u_{2i+1})W_{N/2}^i(y_{N/2}^{N-1}, u_0^{2i-1}|u_{2i+1}),$$

(31)

where $u_0^{2i-1}$ and $u_0^{2i-1}$ denote the vectors formed by extracting the elements of the vector $u_0^{2i-1}$ with even and odd indices, respectively. It should be noted that the equalities $u_0^{2i-1} + u_0^{2i-1} =
\((u_0^{N-1}(I_{N/2} \otimes F)R_N)i_{0}^{i-1} \text{ and } u_0^{2i-1} = (u_0^{N-1}(I_{N/2} \otimes F)R_N)^{N/2+i-1}\) hold, which clearly indicate the relation between the above recursive procedure of channel generation and the recursive structure (13) of the generator matrices \(\{G_N\}\). Successive cancellation decoding is defined as a method of estimating \(u_i\) sequentially via the maximum-likelihood estimation for the channel \(W_N^i\), as

\[\hat{u}_i = \arg \max_{u_i=0,1} W_N^i(y_0^{N-1}, u_0^{i-1}|u_i).\] (32)

The factor graph regarding estimation of \(u_i\) has a tree structure for each \(i\), so that the estimation can be performed efficiently via belief propagation without iteration.

As in the case with binary erasure channels, one can construct polar codes if one knows the probabilities \(P(A_{i,N})\) of incorrectly estimating \(u_i\) via successive cancellation decoding, conditional on the event that one knows the correct estimation results for \(u_0^{i-1}\). Moreover, one can argue theoretical performance of polar codes if one can evaluate how many indices \(i\) there are for which \(P(A_{i,N})\) take values close to 0. We have discussed, in the argument of channel polarization in section 3.3, properties of the stochastic process \(\{p_n\}\) corresponding to the error probabilities \(P(A_{i,N})\), where we had to confirm the following three things in order for the combination of polar codes and successive cancellation decoding to achieve channel capacity: First, to show that \(p_n\) converges to a random variable \(p_\infty\) as \(n \to \infty\). Second, to show that \(p_\infty\) takes either 0 or 1. And third, to show that the probability that \(p_n\) takes values close to 0 asymptotically approaches the channel capacity. It is not an easy task, however, to confirm the above three points when one considers the error probabilities \(P(A_{i,N})\) directly. Arikam [1] considered symmetric channel capacity

\[I(W) = \sum_{x=0,1} \sum_{y} \frac{1}{2} W(y|x) \log \frac{W(y|x)}{\frac{1}{2} [W(y|0) + W(y|1)]}.\] (33)

which is defined for a general binary-input memoryless channel \(W(y|x)\), as well as Bhattacharyya parameter

\[Z(W) = \sum_{y} \sqrt{W(y|0)W(y|1)}.\] (34)

The Bhattacharyya parameter is known to give an upper bound of maximum-likelihood decoding for the channel \(W\). If the channel is a binary erasure channel, then the symmetric capacity and the Bhattacharyya parameter have simple relationship with the erasure probability \(p\) as \(I(W) = 1 - p\) and \(Z(W) = p\). For a general binary-input memoryless channel, it is known that the relations

\[I(W) + Z(W) \geq 1\] (35)
\[I(W)^2 + Z(W)^2 \leq 1\] (36)

hold between \(I(W)\) and \(Z(W)\), and the equality in (35) holds if and only if \(W\) is a binary erasure channel.

On the basis of the series of i.i.d. Bernoulli random variables \(B_1, B_2, \ldots\), each with success probability 1/2, we define a random variable \(i \in \{0, 1, \ldots, N-1\}\) for \(N = 2^q\) whose binary expansion is \(B_nB_{n-1} \cdots B_2B_1\). The stochastic processes \(\{I_n\}\) and \(\{Z_n\}\) are defined therefrom, as

\[I_n = I(W_N^i),\]
\[Z_n = Z(W_N^i).\] (37)
In the argument of channel polarization for a binary erasure channel, it was sufficient to consider properties of the stochastic process \( \{p_n\} \). However, in the argument of channel polarization for an arbitrary binary-input memoryless channel, one has to consider two stochastic processes \( \{I_n\} \) and \( \{Z_n\} \), as described in the following. The stochastic process \( \{I_n\} \) is a martingale because the condition

\[
I(W_{N}^{i}) + I(W_{N}^{i+1}) = 2I(W_{N/2}^{i})
\]

holds. On the other hand, \( \{Z_n\} \) is in general a supermartingale because the inequality

\[
Z(W_{N}^{2i}) + Z(W_{N}^{2i+1}) \leq 2Z(W_{N/2}^{i})
\]

holds. Arıkan [1] showed that the stochastic process \( \{Z_n\} \) converges almost surely to a random variable \( Z_\infty \) which takes 0 or 1, on the basis of an argument similar to that discussed in section 2. Using the fact that the bounded martingale \( \{I_n\} \) also converges almost surely to a random variable \( I_\infty \), as well as the relations (35) and (36), he then showed that \( I_\infty \) also takes 0 or 1 with probability

\[
P(I_\infty) = \begin{cases} 
I(W) & (I_\infty = 0) \\
1 - I(W) & (I_\infty = 1) 
\end{cases}
\]

(40)

The above discussion assures that the number of indices \( i \) for which \( P(A_{i, N}) \) take values close to 0 is approximately \( N I(W) \) as \( N \) becomes sufficiently large. In order to show that polar codes constructed by collecting those indices to form \( T \) achieve the symmetric capacity asymptotically, one has to prove that the probability of the event \( Z_n = o(1/N) \) approaches \( I(W) \). This statement was proved by Arıkan [1]. A proof can be obtained, for example, by applying an argument similar to that given in section 3.3 to the stochastic process \( \{Z_n\} \). We have obtained the following result [3, 4], refining the result of Arıkan and Telatar [2] to include dependence on the code rate.

**Theorem 1** Let \( I(W) \) be the symmetric capacity of a binary-input memoryless channel \( W \). Then the error probability of a polar code with codeword length \( N = 2^n \) and code rate \( R < I(W) \) via successive cancellation decoding satisfies

\[
P(\mathcal{E}_N) = o\left(2^{-2^{(n+4t)^3/2}}\right)
\]

(41)

for an arbitrary \( t \) satisfying \( t < Q^{-1}(R/I(W)) \).

A detailed derivation will be given elsewhere [4]. Hassani and Urbanke [5] independently obtained the same result. It should also be worth noting that Korada et al. [6] have shown that (40) can be directly derived without recourse to discussion of the stochastic process \( \{Z_n\} \) for a binary-input memoryless symmetric channel.

### 3.5. Construction of codes for general binary-input memoryless channel

It is not an easy task to evaluate error probabilities \( \{P(A_{i, N})\} \) efficiently for a general binary-input memoryless channel. Although Arıkan [1] proposed a method that is based on approximate evaluation of Bhattacharyya parameters, it is no longer guaranteed for resulting codes to have the capacity-achieving property. For a general binary-input memoryless symmetric channel, we have proposed use of density evolution to evaluate \( \{P(A_{i, N})\} \), whose computational complexity, in terms of the number of convolution of densities, scales as \( O(N) \) [7]. The use of density evolution is justified because decoding of a component is regarded as belief propagation without cycles. We also showed that the proposed method produces codes that are better in error probability than those obtained by Arıkan’s approximate method [8].
4. Extensions

4.1. Construction using $\ell \times \ell$ matrix

Although polar codes are provably capacity-achieving, performance of polar codes with finite codelength is empirically not so remarkable, if compared with that of LDPC codes with the same codelength. Rather poor performance of polar codes may be ascribed to the code construction and/or the decoding, so that there are several attempts to improve performance of polar codes by modifying the code construction method and/or the decoding algorithm.

One can consider an extension in which an $\ell \times \ell$ matrix, with $\ell \geq 3$, is used in place of the $2 \times 2$ matrix $F$ (11) to construct polar codes. For the case with a binary-input memoryless symmetric channel, Korada et al. [6] obtained the following results. First, they showed that a necessary and sufficient condition for an $\ell \times \ell$ matrix $F$ to induce channel polarization via recursive construction similar to that discussed in section 3.1 is that one cannot make the matrix $F$ upper-triangular by any permutation of columns of $F$. Second, defining the “error exponent” of a matrix $F$ to be the supremum of $\beta$ such that the error probability of successive cancellation decoding for codes of codelength $\ell^n$ is $o\left(2^{-\ell^n \beta}\right)$, they showed that the error exponent of $F$ is given by

$$\frac{1}{\ell} \sum_{k=1}^{\ell} \log \ell D_k,$$

(42)

where $\{D_1, \ldots, D_\ell\}$ are called partial distances of the matrix $F$, defined in [6]. On the basis of this result, they succeeded in obtaining upper and lower bounds of the best possible error exponents, and proved that error exponents can be made as close to 1 as possible by taking $\ell \to \infty$.

Theorem 1 can be extended to the case with an $\ell \times \ell$ matrix. Let $\{B_n\}$ be a sequence of i.i.d. random variables with $P(B_n = k) = 1/\ell$ for all $k \in \{1, \ldots, \ell\}$, and let $\mathbb{E}(F)$ and $\mathbb{V}(F)$ be the mean and variance of the random variables $\log \ell D_{B_n}$. Applying the same argument as that in section 3 to the above random variables, one can prove that the error probability of successive cancellation decoding of polar codes, of codelength $N = \ell^n$, satisfies

$$P(\mathcal{E}_N) = o\left(2^{-n \mathbb{E}(F) + t \sqrt{n \mathbb{V}(F)}}\right),$$

(43)

for any $t$ satisfying $Q(t) > R/I(W)$.

4.2. Channel with $q$-ary input

It is natural to consider channel polarization with a $q$-ary input memoryless channel with $q \geq 3$, rather than a binary-input channel. Şaşoğlu et al. [9] discussed the extension of channel polarization to channels with $q$-ary input, on the basis of the $2 \times 2$ matrix $F$ given in (11), which is regarded as representing a linear transform over the ring $\mathbb{Z}/q\mathbb{Z}$ of residue classes modulo $q$. They proved the following theorem:

**Theorem 2** If $q$ is a prime number, $q$-ary polar codes constructed with $F$ as defined in (11) and successive cancellation decoding satisfies

$$P(\mathcal{E}_N) = o\left(2^{-2n\beta}\right)$$

(44)

for any $\beta < 1/2$.

If $q$ is not a prime number, there are pathological cases in which channel polarization does not occur. In order to circumvent such difficulties, Şaşoğlu et al. [9] proposed introducing
randomness in construction of polar codes, and proved the capacity-achieving property of the resulting codes.

Here, it seems more promising to consider fields rather than rings. We have given some sufficient conditions for an \( \ell \)-input transformation (which can be nonlinear) on a set of \( q \) alphabets to induce channel polarization. An upper bound of error probability is also given, on the basis of which we have found that generator matrices of (generalized) Reed-Solomon codes can be good choices for providing larger error exponents. As an example, a \( 4 \times 4 \) matrix over GF\( (q = 2^2) \) composed of a generator matrix of a generalized Reed-Solomon code over GF\( (2^2) \) has been shown to have the error exponent \( \log \frac{24}{(4 \log 4)} \approx 0.57312 \), which is larger than the error exponent \( 0.51828 \) for the best possible \( \ell \times \ell \) matrix over GF\( (2) \) for all \( \ell \leq 16 \), which was reported in [6]. Detailed arguments of the above results will be presented elsewhere [10].

5. Conclusion
We have reviewed channel polarization and polar codes, putting emphasis on our recent results. In this article, we have not mentioned several other important subjects related to channel polarization and polar codes. They include applications of polar codes to various problems of source coding, as well as exploration of more efficient decoding algorithms other than successive cancellation decoding, for which interested readers are encouraged to refer to, e.g., [11].

Acknowledgments
The author would like to thank Mr. Ryuhei Mori, Kyoto University, for his helpful comments. Support from the Grant-in-Aid for Scientific Research on Priority Areas (No. 18079010) is acknowledged.

References
[1] Arikan E 2009 IEEE Trans. Info. Theory, 55 3051–73
[2] Arikan E and Telatar E 2009 Proc. 2009 Int. Symp. Information Theory (Seoul, Korea) pp 1493–5
[3] Mori R and Tanaka T 2009 Proc. 32nd Symp. Information Theory and its Applications (SITA2009) W43-2 (in Japanese).
[4] Tanaka T and Mori R 2010 Refined rate of channel polarization Preprint arXiv:1001.2067v1 [cs.IT]
[5] Hamed Hassani S and Urbanke R 2010 On the scaling of polar codes: I. The behavior of polarized channels Preprint arXiv:1001.2766v3 [cs.IT]
[6] Korada S B, Şasoglu E, and Urbanke R 2009 Polar codes: Characterization of exponent, bounds, and constructions Preprint arXiv:0908.0302v1 [cs.IT]
[7] Mori R and Tanaka T 2009 Proc. 2009 Int. Symp. Information Theory (Seoul, Korea) pp. 1496–1500
[8] Mori R and Tanaka T 2009 IEEE Comm. Lett., 13 pp. 519–21
[9] Şasoglu E, Telatar E, and Arikan E 2009 Polarization for arbitrary discrete memoryless channels Preprint arXiv:0908.0302v1 [cs.IT]
[10] Mori R and Tanaka T 2010 Channel polarization on q-ary discrete memoryless channels by arbitrary kernels Preprint arXiv:1001.2662v2 [cs.IT]
[11] Korada S B 2009 Polar Codes for Channel and Source Coding Doctoral Thesis, Ecole Polytechnique Fédérale de Lausanne