Universal Polar Coding for Asymmetric Channels

Karthik Nagarjuna and Paul H. Siegel
CMRR, University of California, San Diego
Email:{tkarthik,psiegel}@ucsd.edu

***Abstract*—We present a universal coding scheme, based on polar codes, that can achieve the compound capacity of any finite set of binary-input asymmetric channels. The scheme is a hybrid combination of Honda and Yamamoto's polar coding scheme for asymmetric channels and a universal polar coding scheme for symmetric channels proposed by Hassani and Urbanke. In the proposed universal construction for the asymmetric setting, we exploit the staircase structure in the universal scheme for symmetric channels to define a coding strategy that requires neither storage-intensive shared boolean functions nor a side-channel between encoder and decoder in order to transmit bits corresponding to bit-channels that are not completely polarized.**

I. Introduction

Arikan [1] constructed capacity-achieving codes for binary input symmetric channels. A capacity-achieving coding scheme based on source and channel polarization for binary input asymmetric channels was proposed by Honda and Yamamoto [7] which, following Mondelli et al. [8], we refer to as the integrated scheme. In this scheme, Boolean functions are shared between encoder and decoder for non-information carrying bit-channels, requiring exponential storage complexity. En Gad et al. [4] used randomized rounding for low entropy and not completely polarized bit-channels. In addition, a side channel was used to reliably transmit bits corresponding to not completely polarized bit-channels whose fraction is vanishing with respect to the block length. This reduces the storage requirement to polynomial complexity. It was noted in [8] that better simulation results were achieved when an argmax rule was used in place of randomized rounding to encode low-entropy bits, an observation subsequently confirmed analytically by Chou and Bloch [2].

A compound channel is a set of DMCs, $(\mathcal{X}, \{p_s(y|x) : s \in S\}, \mathcal{Y})$ where $y \in \mathcal{Y}$ for every state s in the set $\mathcal{S}$. The compound channel can be looked at as a DMC with state, where the state is arbitrarily selected and fixed for the transmission of an entire block. The assumption is that the decoder knows the channel state. Hassani and Urbanke [5], [6] presented two "polar-like" universal coding schemes to achieve rates close to the compound capacity of binary-input symmetric discrete-memoryless channels (DMCs). In this paper, we present a universal polar coding scheme for the *asymmetric* setting that combines elements of the integrated scheme in [7] and the staircase construction in [5], [6] when the compound capacity-achieving distribution is non-uniform. Our modifications to these constructions eliminate the need for using either storage-intensive shared Boolean functions or a separate side channel to transmit bits corresponding to bit-channels that are not completely polarized.

In Section II we introduce some notation and recall some background results. In Section III we reformulate the integrated scheme in [7], [4] for achieving the capacity of binary-input asymmetric channels. In Section IV we describe the universal polar coding scheme for binary-input asymmetric DMCs that uses a modification of the staircase construction in [5] to simplify the encoding and decoding of incompletely polarized bit-channels.

II. Preliminaries

In this paper we consider only binary-input asymmetric DMCs. We denote the input alphabet by $\mathcal{X} = \{0, 1\}$ and the output alphabet by $\mathcal{Y}$. We express any set of random variables $X_i, X_{i+1}, ..., X_j$ $(i < j)$ by a row vector $(X_i, X_{i+1}, ..., X_j)$ which is denoted by $X^{i:j}$. We denote the set $\{1, 2, 3, ..., N\}$ by $[N]$. Let $U^{1:N}$ be a row vector and let $\mathcal{A} \subset [N]$. The row vector consisting of elements in $U^{1:N}$ corresponding to the subset of positions $\mathcal{A}$ in the same order from left to right is denoted by $U^{\mathcal{A}}$. We use the abbreviation "w.p." for "with probability".

Let $S = \{1, 2, ..., s\}$, $s \in \mathbb{N}$, be a set of DMCs which are characterized by the conditional distributions $p_l(y|x)$, $l \in S$. Let $(X_1, Y_{l1}), (X_2, Y_{l2}), ..., (X_N, Y_{lN})$ be i.i.d. random tuples distributed according to $P_X(x)p_l(y|x)$, where $l \in S$ and $N = 2^n$. Let G_N be the conventional polar transformation, represented by a binary matrix of dimension $N \times N$. If $U^{1:N} = X^{1:N}G_N$, then we denote $\mathbb{P}(U^{1:N} = u^{1:N})$ by $P_{U^{1:N}}(u^{1:N})$ and similarly we denote $\mathbb{P}(U_i = u_i|U^{1:i-1}Y_l^{1:N} = u^{1:i-1}y_l^{1:N})$ by $P_{U_i|U^{1:i-1}Y_l^{1:N}}(u_i|u^{1:i-1}y_l^{1:N})$. For two random variables (X, Y_l) distributed as $P_X(x)p_l(y|x)$, the Bhattacharya parameter is defined as

$$Z(X|Y_l) = 2\sum_y P_Y(y)\sqrt{P_{X|Y_l}(1|y)P_{X|Y_l}(0|y)}.$$

Let $\beta < 0.5$ and define the following subsets obtained by polarization, with notation adapted from [4].

$$\mathcal{H}_X = \{i \in [N] : Z(U_i|U^{1:(i-1)}) \geq 1 - 2^{-N^\beta}\}.$$
$$\mathcal{L}_X = \{i \in [N] : Z(U_i|U^{1:(i-1)}) \leq 2^{-N^\beta}\}.$$
$$\mathcal{H}_{X|Y_l} = \{i \in [N] : Z(U_i|U^{1:(i-1)}Y_l^{1:N}) \geq 1 - 2^{-N^\beta}\}.$$
$$\mathcal{L}_{X|Y_l} = \{i \in [N] : Z(U_i|U^{1:(i-1)}Y_l^{1:N}) \leq 2^{-N^\beta}\}.$$

Note that $\mathcal{L}_X \subseteq \mathcal{L}_{X|Y_l}$, $l \in S$. From Theorem 1 in [7] we have the following results.

$$\lim_{N\to\infty} \frac{1}{N}|\mathcal{H}_X| = H(X).$$
$$\lim_{N\to\infty} \frac{1}{N}|\mathcal{L}_X| = 1 - H(X).$$
$$\lim_{N\to\infty} \frac{1}{N}|\mathcal{H}_{X|Y_l}| = H(X|Y_l).$$
$$\lim_{N\to\infty} \frac{1}{N}|\mathcal{L}_{X|Y_l}| = 1 - H(X|Y_l).$$

We define several other subsets of bit-channels as follows.

$$I_l = \mathcal{H}_X \cap \mathcal{L}_{X|Y_l}$$
$$F_l = \mathcal{H}_X \cap \mathcal{L}^c_{X|Y_l}$$
$$D = \mathcal{L}_X$$
$$R = (\mathcal{H}_X \cup \mathcal{L}_X)^c$$

We refer to these as good, bad, deterministic, and not completely polarized bit-channels, respectively.

The capacity of a compound channel is well known [3, p. 170] and is given by

$$C_c = \max_{P_X(x)} \min_{l\in S} I(X; Y_l).$$

where (X, Y_l) is distributed as $P_X(x)p_l(y|x)$.

III. Polar code for asymmetric channels

In this section, we present the capacity-achieving asymmetric channel coding scheme based upon [7], [4], [2] which is used as a building block in our proposed universal polar coding scheme. Let the asymmetric DMC be characterized by $p(y|x)$ and let $p(x)$ be the non-uniform capacity-achieving input distribution. We use the same notation as in Section II with the substitution of Y, I and F for Y_l, I_l and F_l respectively as we are considering here the single channel case. Now we describe the encoding and decoding procedure.

A. Code construction

We first generate random function $f : F \to \{0, 1\}$, where each $f(j)$, $j \in F$ is chosen independently and uniformly. These frozen bits are shared between encoder and decoder.

We also generate independent random boolean functions $\lambda_i : \{0, 1\}^{i-1} \to \{0, 1\}$ for each $i \in R$ by using the following probability rule:

$$\lambda_i(u^{1:i-1}) = u \text{ w.p. } P_{U_i|U^{1:i-1}}(u|u^{1:i-1}), \text{ for } u \in \{0, 1\}$$

independently for each $u^{1:i-1}$. Let the set of random functions be denoted by λ_R. These functions are shared between encoder and decoder, which can require exponential storage complexity.

Encoding
Input: randomly chosen message $M^{1:|I|}$
Output: codeword $X^{1:N}$
for $i = 1 : N$, encode U_i as follows.
1. If $i \in I$, the value of U_i is given by setting $U^I = M^{1:|I|}$.
2. If $i \in F$, we set $U_i = f(i)$.
3. If $i \in D$, we encode U_i using the **argmax rule** [2]
$$U_i = \text{argmax}_{x\in\{0,1\}} P_{U_i|U^{1:i-1}}(x|U^{1:i-1}).$$
4. If $i \in R$, we set $U_i = \lambda_i(U^{1:i-1})$.
end
Transmit $X^{1:N} = U^{1:N}G_N$.

The decoding algorithm is as follows.

Decoding
Input: received vector $Y^{1:N}$
Output: message estimate $\hat{M}^{1:|I|}$
for $i = 1 : N$
1. If $i \in F$, set $\hat{U}_i = f(i)$
2. If $i \in \mathcal{L}_X \cup I$, set
$\hat{U}_i = \mathbb{1}\{P_{U_i|U^{1:i-1},Y^{1:N}}(1|\hat{U}^{1:i-1}, Y^{1:N}) \geq P_{U_i|U^{1:i-1},Y^{1:N}}(0|\hat{U}^{1:i-1}, Y^{1:N})\}$
3. If $i \in R$, set $\hat{U}_i = \lambda_i(\hat{U}^{1:i-1})$.
end
Decode $\hat{M} = \hat{U}^I$.

For $i \in D$, the induced conditional distribution $\delta_i(u|u^{1:i-1})$ on U_i given $U^{1:i-1}$ satisfies $\delta_i(u|u^{1:i-1}) = 1$ and $\delta_i(u + 1|u^{1:i-1}) = 0$ where

$$u = \text{argmax}_{x\in\{0,1\}} P_{U_i|U^{1:i-1}}(x|u^{1:i-1}).$$

The ensemble average distribution of $U^{1:N}$ is

$$\mathbb{E}_{(\lambda_R,f)}[\mathbb{P}(U^{1:N} = u^{1:N}|(\lambda_R, f))]$$
$$= 2^{-|\mathcal{H}_X|}\Pi_{i\in R} P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1})\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1}).$$

This average distribution is $O(2^{-N^{\beta'}})$ close in total variation distance to the distribution when $X^{1:N}$ is an i.i.d. random vector, for $\beta' < 0.5$. Therefore, the decoding method is reliable, with average probability of error $\mathbb{E}_{(\lambda_R,f)}[P_e(\lambda_R, f)] = O(2^{-N^{\beta'}})$ [7]. In [4], use of a side-channel is proposed for bit-channels $(\mathcal{H}_X \cup \mathcal{L}_{X|Y})^c$ as an alternative to sharing boolean functions. In contrast to [4] we propose sharing boolean functions for bit-channels $(\mathcal{H}_X \cup \mathcal{L}_X)^c$ to enable the use of this scheme for universal coding. The quantities $P_{U_i|U^{1:i-1}}(u|u^{1:i-1})$ and $P_{U_i|U^{1:i-1},Y^{1:N}}(u|u^{1:i-1}, y^{1:N})$ used during encoding and decoding can be computed in $O(N \log N)$ real operations using techniques in [7].

IV. Universal Scheme for Asymmetric Channels

Let $p(x)$ be the non-uniform compound capacity-achieving distribution for compound channel S. Consider the good bit-channels I_l and bad bit-channels F_l for $l \in S$, as well as the deterministic channels D and not competely polarized channels R. Note that $|R|$ is a vanishing fraction of the block length as N increases.

Let $L = \min\{|I_1|, |I_2|..., |I_s|\}$. Clearly, $\lim_{N\to\infty} \frac{L}{N} = C_c$. If the inequality $|I_1 \cap I_2 \cap ... \cap I_s| \leq L$ is strict, then not all channels in the set S share the same set of good bit-channels.

By assigning message bits to indices in $I_1 \cap I_2 \cap ... \cap I_s$, uniform random frozen bits to indices in $\mathcal{H}_X - (I_1 \cap I_2 \cap ... \cap I_s)$, and encoded bits derived using the same coding scheme as in Section III, we can get a reliable code, but it will not be capacity-achieving.

Note that $F_l \cup I_l = \mathcal{H}_X$ for all $l \in S$. This implies that for any channels $l, m \in S$, $l \neq m$, a bit-channel which is good for channel l and not good for channel m will be a bad bit-channel for channel m. This fact will enable us to adapt the universal coding scheme for symmetric channels [5] to the asymmetric case and to construct codes that achieve rates close to $\frac{L}{N}$.

As in [5], [6], we use a staircase composed of polar blocks to achieve rates close to $\frac{L}{N}$ and universality. We exploit the staircase structure to avoid sharing storage-intensive boolean functions bits corresponding to not completely polarized indices in R. To do so, we initially assume $|I_1 \cap I_2 \cap .... \cap I_s| \geq |R|$, an assumption that will be relaxed later (see Remark IV-B.2). Take I' to be a subset of $I_1 \cap I_2 \cap .... \cap I_s$ such that $|I'| = |R|$. Define an arbitrary bijection $g : I' \to R$. Let $L' = \min\{|I_1|, |I_2|, ...|I_s|\} - |I'|$.

A. Code construction

Generate a random frozen vector $W^{1:N}$ such that

$$\mathbb{P}(W^{1:N} = u^{1:N}) = 2^{-(|\mathcal{H}_X|)} \Pi_{i \in R} P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}) \Pi_{i \in \mathcal{L}_X} \delta_i(u_i|u^{1:i-1}).$$

The vector $W^{1:N}$ is shared between encoder and decoder. Let us assume that there is a linear MDS code $\mathcal{M}$ with blocklength $|\mathcal{H}_X| - |R|$ over a field $GF(2^q)$ for some $q \in \mathbb{N}$, and consider its equivalent binary linear code representation. We will make us of the following result to achieve universality in our construction.

Lemma 1. *Let G be the generator matrix of the linear MDS code $\mathcal{M}$ over $GF(2^q)$. If G does not have a zero column, then any position in the binary representation of codewords of $\mathcal{M}$ will have an equal number of zeros and ones.*

Proof: Let j be any column of G. Since it is non-zero, it has a non-zero entry $g_{ij} \in GF(2^q)$. The jth position of the codeword corresponding to message $[0, \ldots, m_i, \ldots, 0]$ will be $m_i g_{ij}$. A m_i ranges over all elements of $GF(2^q)$, $m_i g_{ij}$ also does. Therefore the binary representation of this codeword entry ranges over all possible binary q-tuples. This ensures that for any position in the binary representation of $\mathcal{M}$ there exists a codeword which has the value 1 in that position. Due to linearity of the equivalent binary representation, we must have an equal number of zeros and ones in that position.

We arrange polar blocks of size N, for N sufficiently large, in a staircase with height N. We extend the staircase by placing $k \in \mathbb{N}$ such staircases side-by-side. Now take q such extended staircases, graphically placed one above the other, as illustrated in Fig. 1 for the case $N = 6$, $k = 3$, and $q = 2$. While encoding we fill all the polar blocks column-by-column from left to right in the staircase structure, and we follow the same order for decoding. Hence we encode/decode

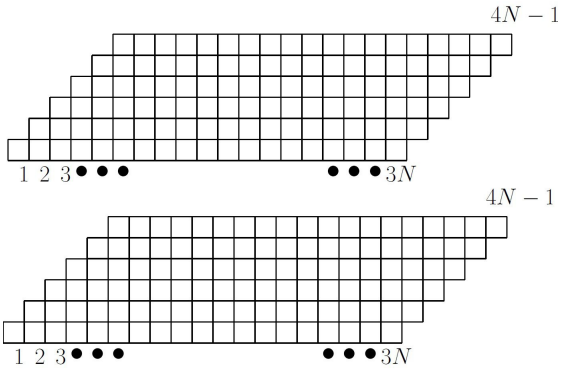

Fig. 1. Extended staircases with k=3, N=6 and q =2

different polar blocks in parallel while encoding/decoding a column. The total number of columns is $(k+1)N - 1$, and we label them with indices $1 : (k+1)N - 1$ from left to right. Our goal is to encode each polar block in the staircase with the same ensemble average distribution as we produced in the asymmetric channel coding so that decoding will be reliable. The encoding and decoding schemes are as follows.

Encoding
Input: qL' information bits for each full column.
Output: $U^{1:N}$ of each polar block in the staircase.

- To encode non-full-height columns on the left from $t = 1 : N-1$, we assign $U_i = W_i$ for the block with channel index i in that column. Repeat this for all q staircases. This step ensures that the prefix part of the polar blocks satisfies the required ensemble average distribution.
- To encode full-height columns from $t = N \leq i \leq kN$:
  - First, encode the blocks with index $i \in \mathcal{L}_X$ in column t using the argmax rule. Repeat this for all q staircases. This maintains the required conditional distribution for these indices.
  - Second, encode the blocks with index $i \in R$ in column t using the randomized rounding rule, i.e., $U_i = u$ w.p. $P_{U_i|U^{1:i-1}}(u|U^{1:i-1})$ for $d \in \{0, 1\}$. Repeat this for all q staircases. This will maintain the required conditional distribution. Since these are randomly generated, we use the inverse function g^{-1} to copy these bits to $I' \subseteq I_1 \cap I_2 \cap ... \cap I_s$ where they can be reliably decoded.
  - Third, finish encoding the blocks with index $i \in I'$ by assigning $U_i = X \oplus U'_{g(i)}$, where $U'_{g(i)}$ is the bit copied from the block with index $g(i) \in R$ and X is the parity of the information bits corresponding to that column. We do the same for all q staircases. This maintains the distribution of the indices in $\mathcal{H}_X$ and also ensures the independence from previously encoded bits of the polar block. This is the key step of the construction, since direct use of $U'_{g(i)}$ to encode U_i would not satisfy the required distribution.
  - Fourth, encode the blocks with indices $i \in \mathcal{H}_X - I'$.

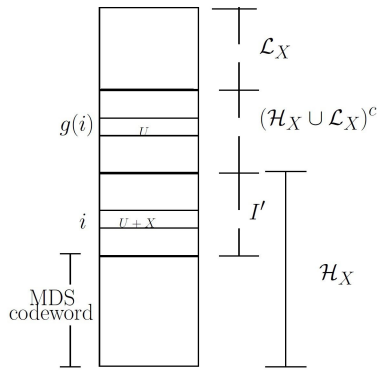

Fig. 2. Full-height column structure: X is the parity of the information bits of the column and U is the encoded bit in the block with index $g(i)$ in the column.

  - $*$ Encode qL' information bits (equivalent to L' symbols over $GF(2^q)$) into codeword m in the binary representation of $\mathcal{M}$.
  - $*$ Fill blocks with indices in $i \in \mathcal{H}_X - I'$ in all q staircases with codeword m as shown in Fig. 2. By Lemma 1, a uniform distribution is guaranteed for these positions, as required for indices in $\mathcal{H}_X$. Since m depends only on the information bits of the current column, independence from previously encoded bits of the polar block is also guaranteed.
- To encode non-full-height columns $t = kN + 1 : (k + 1)N - 1$ on the right, we generate all bits randomly to satisfy the distribution of the polar block. This is done as follows:
  - For blocks with index $i \in \mathcal{H}_X$ generate U_i independently and uniformly.
  - For blocks with index $i \in R$, generate $U_i = u$ w.p. $P_{U_i|U^{1:i-1}}(u|U^{1:i-1})$, for $u \in \{0, 1\}$.
  - For blocks with index $i \in \mathcal{L}_X$ use argmax rule.
- Transmit $X^{1:N} = U^{1:N}G_N$ for each polar block.

Decoding

Input: Received vector $Y^{1:N}$ for each block.

Output: Estimates of encoded information bits.

- To decode non-full-height columns on the left from $t = 1 : N - 1$, we estimate $\hat{U}_i = W_i$ for the block with channel index i in that column. Repeat this for all q staircases.
- To decode full-height columns from $t = N \leq i \leq kN$:
  - First, decode the blocks with index $i \in \mathcal{L}_X \cup I'$ in column t using standard successive cancellation decoding. This is possible since these indices are either good for all channels or deterministic.
  - Second, decode the blocks with index in $\mathcal{H}_X - I'$:
    - $*$ Decode the L' symbols from the good indices based on the channel that is selected. Let C be the partially recovered codeword.
    - $*$ The codeword $\hat{m}$ can be recovered from c since it is an MDS codeword, providing an estimate of qL' information bits corresponding to the column. Let $\hat{X}$ be the parity of the decoded information bits in the column.
  - Last, decode blocks with index $i \in R$ by estimating $\hat{U}_i = \hat{X} \oplus \hat{U}'_{g^{-1}(i)}$ where $\hat{U}'_{g^{-1}(i)}$ is the already decoded bit corresponding to the block with index $g^{-1}(i) \in I'$ ih the same column.
- Ignore and do not decode non-full-height columns $t = kN + 1 : (k + 1)N - 1$ on the right. Note that this will not block further decoding.

Note that we encoded $L'q$ information bits only in full-height columns. Hence we get the rate $\frac{L'}{N}$ for each full-height column. Since $\frac{|I'|}{N}$ is diminishing, the rate for each such column will be close to $\frac{L}{N}$. Also, as k increases, only full-height columns will constitute a significant fraction of the code bits. The exact relation between achievable rate R and k can be found in [5], [6].

Now we derive an upper bound on q, which also upper bounds the total number of polar blocks in the staircase structure. If we consider a Reed-Solomon (RS) code as the linear MDS code over $GF(2^q)$, the blocklength of the code should divide $2^q - 1$. We bound q as follows.

- If $|\mathcal{H}_X| - |R|$ is odd:
  By Euler's Theorem, q can take value $\phi(|\mathcal{H}_X| - |R|)$ where ϕ is Euler's totient function. Therefore $q \leq \phi(|\mathcal{H}_X| - |R|) \leq |\mathcal{H}_X| - |R| \leq N$.
- If $|\mathcal{H}_X| - |R|$ is even:
  Use a RS code of blocklength $(|\mathcal{H}_X| - |R|) - 1$. Then $q \leq N$ since the blocklength is odd. Fill the remaining position with the parity of the information bits to maintain the distribution of the set $\mathcal{H}_X$ in all q staircases and modify the scheme accordingly.

The description of the coding scheme indicates why each polar block will satisfy the required distribution. The following theorem charaterizes the probabiity of error and encoding/decoding complexity of the scheme.

Theorem 1.

1. For every polar block ($U^{1:N}$) encoded in the staircase $\mathbb{E}_{W^{1:N}}[\mathbb{P}(U^{1:N} = u^{1:N}|W^{1:N})] = 2^{-(|\mathcal{H}_X|)}\Pi_{i\in R}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1})\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1})$.

2. The average probability of error is $\mathbb{E}_{W^{1:N}}[P_e(W^{1:N})] = NqkO(2^{-N^\beta})$ *for* $\beta < 0.5$.

3. Encoding and decoding take $O((\log_2 N)q^{\log_2 3-1})$ *and* $O((\log_2 N)^2 q^{\log_2 3-1})$ *binary operations per bit, respectively. Encoding and decoding also take* $O(\log_2 N)$ *real operations per bit.*

Proof Refer to the Appendix.

B. Remarks

1 We used a linear MDS code to satisfy the distribution for each block. In the symmetric channel construction, linearity is not required.

2 If the condition $|I_1 \cap I_2 \cap .... \cap I_s| \geq |R|$ does not hold, we can use the universalizing technique based on polarizing

indices [5] with the required number of recursions to produce a partially universalized block that satisfies the desired condition. We can then use this partially universalized block in the staircase structure.

V. Conclusion

We presented a universal polar coding scheme for a compound channel defined by a finite set of binary-input asymmetric DMCs with non-uniform compound capacity-achieving input distribution. The proposed scheme exploits the underlying staircase structure in the code construction to avoid the need for either side-channel transmission or storage-intensive Boolean functions for bits corresponding to not completely polarized bit-channels. The scheme requires a large block length, leaving open the problem of designing codes with short block length. Another open problem is the construction of a stronger universal polar code with reduced storage complexity, which achieves rate R less than compound capacity with non-uniform compound capacity-achieving input distribution $p(x)$ and also achieves rate R for any channel whose mutual information evaluated at $p(x)$ is larger than R.

VI. Acknowledgement

This work was supported by National Science Foundation Grants CCF-1415109 and CCF-1619053.

References

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

[2] R. A. Chou and M. R. Bloch "Using deterministic decisions for low-entropy bits in the encoding and decoding of polar codes," *Proc. 53rd Annu. Allerton Conf. on Commun., Control, and Computing (Allerton 2015),* Monticello, IL, Sep. 2015, pp. 1380–1385.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., New York: Wiley, 2006.

[4] E. En Gad, Y. Li, J. Kliewer, M. Langberg, A. A. Jiang and J. Bruck, "Asymmetric error correction and flash-memory rewriting using polar codes," *IEEE Trans. Inf. Theory,* vol. 62, no. 7, pp. 4024–4038, Jul. 2016.

[5] S. H. Hassani and R. L. Urbanke, *Universal polar codes, CoRR (2013),*abs/1307.7223.

[6] S. H. Hassani and R. L. Urbanke, "Universal polar codes," *Proc. IEEE Int. Symp. Inf. Theory,* Honolulu, HI, Jul. 2014, pp. 1451–1455.

[7] J. Honda and H. Yamamoto, "Polar coding without alphabet extension for asymmetric models," *IEEE Trans. Inf. Theory,* vol. 59, no. 12, pp. 7829–7838, Dec. 2013.

[8] M. Mondelli, R. Urbanke and S. H. Hassani, "How to achieve the capacity of asymmetric channels," *Proc. 52th Annu. Allerton Conf. on Commun., Control, and Computing (Allerton 2014),* Monticello, IL, Oct. 2014, pp. 789–796.

[9] V. Taranalli, H. Uchikawa and P. H. Siegel, "On the capacity of the beta-binomial channel model for multi-level cell flash memories," *Proc. IEEE J. Select. Areas Commun.* vol. 34, no. 9, pp. 2312–2324, Sep. 2016.

Appendix A
Proof of Theorem 1

We summarize the important steps of the proof. Details are omitted due to space constraints.

Step 1: Consider any polar block in the extended staircase structure which lies completely in the full-height column region. This polar block corresponds to N distinct full-height columns. To get the distribution on encoded codeword $U^{1:N}$ for such a block, we first compute the conditional distribution $\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N})$ for all i in the block.
If $i \in \mathcal{L}_X$, then by the encoding rule
$\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N}) = \delta_i(u_i|u^{1:i-1})$.
If $i \in R$, then by the encoding rule
$\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N}) = P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1})$.
If $i \in \mathcal{H}_X - I'$. then by Lemma 1
$\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N}) = 0.5$.
If $i \in I'$, then by the encoding rule
$\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N}) = 0.5$.
Now U_i is encoded as $X \oplus Y$ where X is the parity of the information bits in the column which are Bernoulli(0.5) and independent of Y, which, in turn, is Bernoulli(p) for some p, but dependent on the previously encoded columns. It follows that U_i is a Bernoulli(0.5) random variable and independent of previous columns. Together, these properties imply that, for an encoded block belonging to full-height columns, $U^{1:N}$ is distributed as

$$\begin{aligned}&\mathbb{P}(U^{1:N} = u^{1:N}|W^{1:N})\\ &= \Pi_{i\in[N]}\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N}) \qquad (1)\\ &= 2^{-|\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1})\end{aligned}$$

Hence we get the required ensemble average distribution for such blocks.

Step 2: Consider a polar block which is partly in the non-full-height region on the right side. For U_i in a full-height column regime, the conditional probability rule is the same as above. For U_i in a non-full-height column, by the encoding rule we get the required conditional distribution. Therefore, we get the same ensemble average distribution of $U^{1:N}$ as above.

Step 3: Now consider a block which is partly in the non-full-height column region on the left of the staircase structure and partly in the full- height region. Let H be the index set of the non-full-height region, with encoded bit W_i corresponding to the index i. Then

$$\begin{aligned}&\mathbb{E}_{W^{1:N}}[1(\cap_{i\in H}(u_i = W_i))]\\ &\overset{(a)}{=} 2^{-|H\cap\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X\cap H}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R\cap H} \qquad (2)\\ &\qquad P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1})\end{aligned}$$

where (a) follows by marginalizing over the distribution of $W^{1:N}$. For U_i in the full-height columns, the required conditional distribution is maintained as discussed in the earlier steps. By combining everything, we get the required ensemble average distribution of $U^{1:N}$ for these blocks as well. This concludes the proof of part 1. The average probability of error for each polar block is bounded in a similar manner to the proof of Theorem 3 in [7] once we have the required ensemble average distribution. Hence we get part 2 by using the union bound. Encoding and decoding a polar block takes $O(N\log_2 N)$ real operations. Encoding and decoding a RS codeword can be done in $O(L'\log_2 L')$ and $O(L'(\log_2 L')^2)$ operations over GF(2^q) respectively. Addition and multiplication over this field takes q and $q^{\log_2 3}$ binary operations respectively. One RS codeword corresponds to Nq bits in the extended staircase. This concludes part 3.