Universal Polar Coding for Asymmetric Channels

Karthik Nagarjuna Tunuguntla, and Paul H. Siegel, *Life Fellow, IEEE*

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 [5] 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. We need a condition that the cardinality of the intersection of good bit-channels will be greater than the cardinality of not completely polarized bit-channels in our staircase construction. If the condition is not satisfied, we propose to make use of another universal polar coding scheme based on bit-channel combining [5] to produce a hybrid polar block which satisfies the required condition and we use such a block in the staircase scheme. In particular, we present an algorithm based on this scheme to produce a hybrid polar block of length at most 2^{s-1} times the original polar block length for satisfying the desired condition, where s is the number of DMCs in the compound channel.

Keywords: Polar codes, binary-input asymmetric channel, compound channel

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

Karthik Nagarjuna Tunuguntla and Paul H. Siegel are with the Department of Electrical and Computer Engineering and the Center for Memory and Recording Research, University of California, San Diego, La Jolla, CA 92093, U.S.A. (e-mail: {tkarthik, psiegel}@ucsd.edu). This work was presented in part at the IEEE Information Theory Workshop, Guangzhou, China, Nov 25-29, 2018, pp 1-5.

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 discrete-memoryless channels (DMCs), $(\mathcal{X}, \{p_s(y|x) : s \in S\}, \mathcal{Y})$ where $y \in \mathcal{Y}$ for every state s in the set 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 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 for achieving the capacity of binary-input asymmetric channels in [7], [4]. 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. We initially assume a condition on the polar block for the compound channel. If the condition is not satisfied by the polar block, we propose to instead use a hybrid polar block in the staircase that is produced using a universalization technique based on bit-channel combining [5] for satisfying the required condition. In Section V, we propose an algorithm to produce the hybrid polar block at a specified block length for the staircase scheme. In Section VI, we illustrate the encoding and decoding method in the case of compound channel with two states for a block produced after the first step of the recursive procedure in the universalization technique based on bit-channel combining.

II. Preliminaries

We denote the input alphabet of the compound channel by $\mathcal{X} = \{0, 1\}$ and the output alphabet by $\mathcal{Y}$. We express any set of random variables $X_i, X_{i+1}, \ldots, X_j$ $(i < j)$ by a row vector $(X_i, X_{i+1}, \ldots, X_j)$ which is denoted by $X^{i:j}$. We denote the set $\{1, 2, 3, \ldots, N\}$ by $[N]$. We denote the set $\{i, i+1, \cdots, j\}$ by $[i : j]$ $(i < j)$. 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 positions in $\mathcal{A}$ is denoted by $U^{\mathcal{A}}$. We use the abbreviation "w.p." for "with probability".

Let $S = \{1, 2, \ldots, s\}$, $s \in \mathbb{N}$, be a set of DMCs which are characterized by the conditional distributions $p_l(y|x)$, $l \in S$. Let $(X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N)$ be identical and independent (iid) 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 [1], 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 $\mathbb{P}(U_i = u_i|U^{1:i-1}Y^{1:N} = u^{1:i-1}y^{1:N})$ by $P_{U_i|U^{1:i-1}Y^{1:N}}(u_i|u^{1:i-1}y^{1:N})$. For two random variables (X, Y) distributed as $P_X(x)p_l(y|x)$, the Bhattacharya parameter is defined as

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

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

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

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

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

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

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

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

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

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

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

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. 172] and is given by

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

where (X,Y) is distributed as $P_X(x)p_l(y|x)$. The compound capacity-achieving distribution could be non-uniform only if at least one of the DMCs in S is asymmetric. In this paper, we consider universal polar coding with non-uniform input distribution.

Example:

Let $S = \{1,2\}$. Let DMC 1 be a Z-channel with cross-over probabilities $p(0|1) = 0.5$, $p(1|1) = 0.5$ and $p(0|0) = 1$. Let DMC 2 be a binary erasure channel with erasure probability 0.5. Here $I(X;Y_2)$ dominates $I(X;Y_1)$ for all input distributions. Hence the compound capacity-achieving distribution is same as the capacity-achieving distribution of DMC 1. If X is distributed as Bernoulli(α), then mutual information becomes $I(X;Y_1) = H(Y_1) - H(Y_1|X) = H(\frac{\alpha}{2}) - \alpha$. The derivative of mutual information w.r.t α becomes $\frac{d}{d\alpha} I(X;Y_1) = \frac{1}{2}\log(\frac{1-\alpha/2}{\alpha/2}) - 1$. By equating the derivative to zero, we get $\alpha = 2/5$. This gives the capacity-achieving distribution for DMC 1 as mutual information is concave in α. It will also be the compound capacity-achieving distribution for the compound channel S. Therefore the compound capacity of S is $H(1/5) - 2/5 = 0.322$.

III. Polar coding for binary-input asymmetric channels using integrated scheme

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 I and F for 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. The encoding algorithm is described as follows.

```
Encoding
Input: Uniformly distributed 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 $\hat{M}^{1:|I|} = \hat{U}^I$.
```

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. Staircase Scheme with non-uniform input distribution

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 with respect to the block length as N increases.

Let $L = \min\{|I_1|, |I_2|, \ldots, |I_s|\}$. Clearly, $\lim_{N\to\infty}\frac{L}{N} = C_c$. If the inequality $|I_1 \cap I_2 \cap \ldots \cap I_s| \leq L$ 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 $I_1 \cap I_2 \cap \ldots \cap I_s$, uniform random frozen bits to indices in $\mathcal{H}_X - (I_1 \cap I_2 \cap \ldots \cap I_s)$, 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 for encoding 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 in Section V. Take I' to be a subset of $I_1 \cap I_2 \cap .... \cap I_s$ such that $|I'| = |R|$. Define an arbitrary bijection $g : I' \rightarrow R$. Now we are ready to present the rigorous code construction illustrating the encoding and decoding schemes. Let $L' = \min\{|I_1|, |I_2|, \ldots, |I_s|\} - |I'|$.

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 Maximum distance seperable (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 use of the following lemma in our universal code 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 $m_i g_{ij}$. As 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

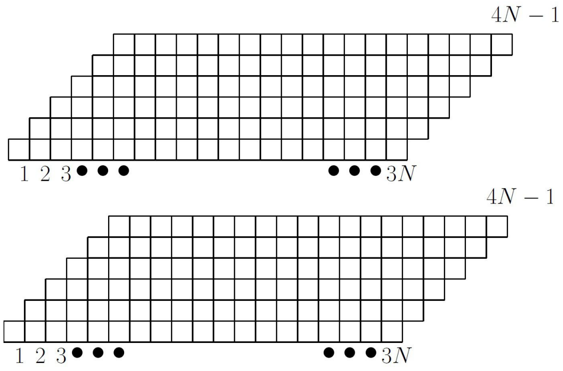

Figure 1. Extended staircases with $k = 3, N = 6$ and $q = 2$

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 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-height 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.

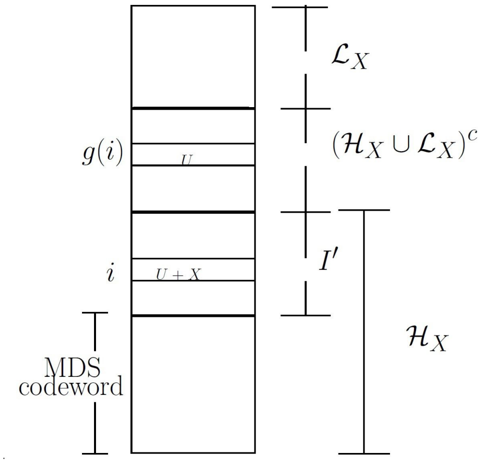

Figure 2. Full-height column structure: X is the parity of the information bits of the column and U is the encoded bit in the block with index $g(i)$ in the column.

This will maintain the required conditional distribution. Since these are randomly generated, we use the inverse function g^{-1} to copy these bits to $I' \subseteq I_1 \cap I_2 \cap \ldots \cap I_s$ 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 the direct use of $U'_{g(i)}$ to encode U_i would not satisfy the required distribution.
- Fourth, encode the blocks with indices $i \in \mathcal{H}_X - I'$.
  - 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 in certain order.

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 by erasure decoding 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 $g^{-1}(i) \in I'$ in 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 prevent 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, the full-height columns will constitute a significant fraction of the total block length. The exact relation between achievable rate R and k can be found in

[5], [6].

We used a linear MDS code in our asymmetric staircase construction. Notice that in the symmetric channel construction, linearity is not required. 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 decoding probability of error and encoding/decoding complexity of the scheme.

Theorem 1.

1. For every polar block 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. Let $P_{e,l}(W^{1:N})$ be the probability of error when DMC l is selected in S for a given code in the above random code construction. The average probability of error is $\mathbb{E}_{W^{1:N}}[P_{e,l}(W^{1:N})] = O(Nqk2^{-N^{\beta}})$ for $\beta < 0.5$ for each $l \in S$.*

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. Existence of universal code with high probability

Theorem 1 states that the average probability of error over the random ensemble is the same for any DMC that gets selected in S. We now show the existence of universal codes with high probability in the random ensemble of codes. Let K be a positive constant. By the Markov inequality, we get the following:

$$\mathbb{P}_{W^{1:N}}(P_{e,l} > Ks\mathbb{E}_{W^{1:N}}[P_{e,l}(W^{1:N})]) < 1/Ks.$$

By the union bound, we get the following.

$$\mathbb{P}_{W^{1:N}}(\cup_{l\in S}(P_{e,l} > Ks\mathbb{E}_{W^{1:N}}[P_{e,l}(W^{1:N})])) < 1/K.$$

Therefore by taking the probability of the complementary event we get

$$\mathbb{P}_{W^{1:N}}(\cap_{l\in S}(P_{e,l} \leq Ks\mathbb{E}_{W^{1:N}}[P_{e,l}(W^{1:N})])) \geq 1 - 1/K.$$

By substituting $K = N$ in the above equation, universal codes that have the probability of error $O(sN^2qk2^{-N^{\beta}})$ exist with high probability $1 - 1/N$ in the random ensemble of codes. Such an analysis is not needed in the symmetric channel case as we have an explicit universal code construction without randomization.

C. Delay advantage in staircase scheme

We have k sub-staircases adjoined to form the extended staircase which gives the total block length. We need to increase the width of the staircase k to infinity for achieving the rates arbitrary close to $\frac{L}{N}$. Such a large k contributes to a very large block length. We can still optimize the delay in communication by overcoming the large width factor k. The idea of saving delay here is due to a possible continuous encoding followed by decoding of sub-staircases sequentially rather than encoding the whole block followed by decoding the whole block. Let us index such sub-staircase components as $m = 1, 2, \cdots, k$.

- Set $j = 1$. We encode U_is for columns $t = 1 : N - 1 + N$. Then we can encode and transmit $X^{1:N}$ corresponding to the polar blocks in sub-staircase $m = 1$. So the decoder has received vectors corresponding to the polar blocks in sub-staircase $m = 1$. Now we decode $\hat{U}_i$s for columns $t = 1 : N$. Increment j.
- Now we encode U_is for next N columns $t = jN : jN + N - 1$. Then we can encode and transmit $X^{1:N}$ corresponding to the polar blocks in sub-staircase $m = j$. So the decoder has received vectors corresponding to the polar blocks in sub-staircase $m = j$. Now we decode $\hat{U}_i$s for columns $t = (j-1)N + 1 : jN$. Decode $\hat{X}^{1:N}$ corresponding to the polar blocks in sub-staircase $j - 1$. Increment j.

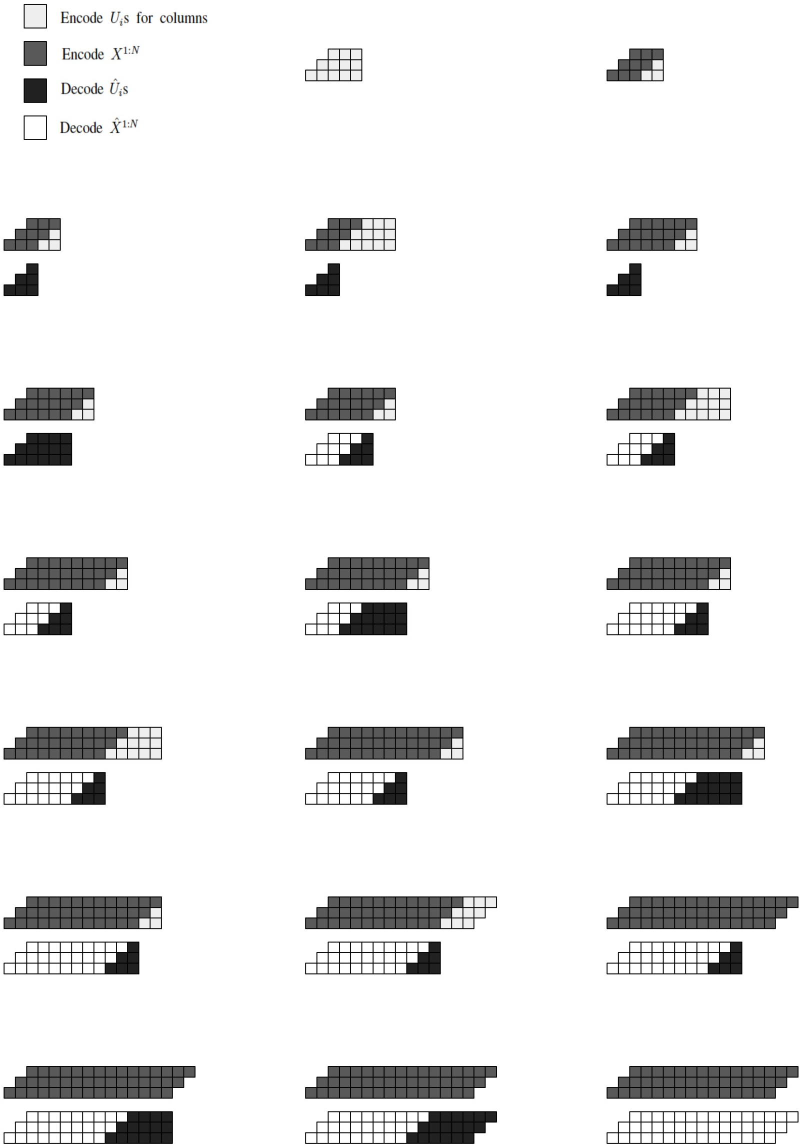

Figure 3. Delay saving through continuous encoding and decoding when $N = 3, k = 5$ and $q = 1$

- We repeat the above step until $j = k$.
- We encode U_is until the last column. Then we can encode and transmit $X^{1:N}$ corresponding to the polar blocks in sub-staircase $m = k$. So the decoder has received vectors correspond-

ing to the polar blocks in sub-staircase $m = k$. Now we can decode $\hat{U}_i$s until the last column.

Now we recover the $\hat{X}^{1:N}$ corresponding to the polar blocks in sub-staircases $k-1$ and k. Figure 3 illustrates these sequential steps pictorially. Notice that this delay saving is general and applies to both symmetric and asymmetric settings.

V. Hybridized Staircase Scheme

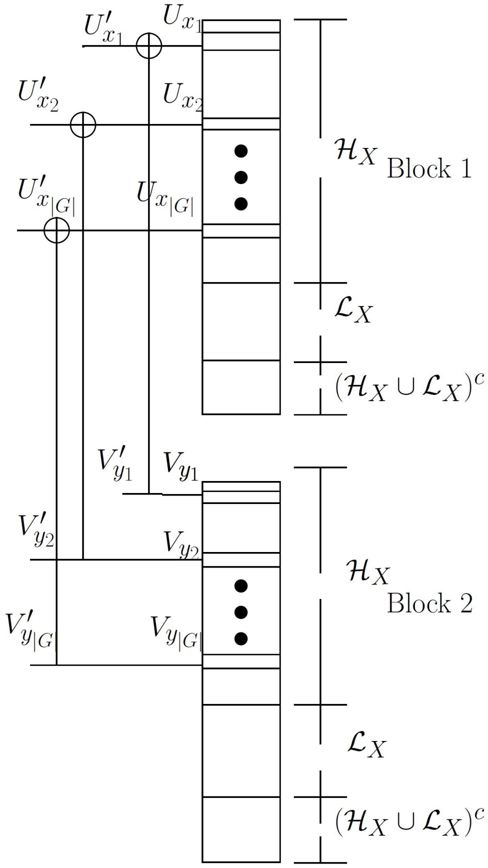

Figure 4. Universal polar code based on combining several bit-channels

If the required condition $|I_1 \cap I_2 \cap \ldots \cap I_s| \geq |R|$ does not hold, we can use the universalizing procedure based on bit-channel combining [5] to produce a partially universalized block that satisfies the desired condition. We propose to use such a hybrid polar block in the staircase structure. Let us briefly discuss the idea of the scheme, which can be readily adapted to the asymmetric case.

The idea of the scheme can be explained by considering two independent polar blocks. If we combine (standard polar combining operation [1]) a bit-channel of the first polar block which is good (resp. bad) for DMCs say $1, 2, \ldots, i$ and bad (resp.good) for DMCs $i+1, i+2, \ldots, s$ with a bit-channel of the second polar block that is bad (resp. good) for DMCs $1, 2, \ldots, i$ and good (resp. bad) for DMCs $i+1, i+2, \ldots, s$, then we get two new bit-channels, one bad for all DMCs $1, 2, \ldots, s$ and the other one good for all DMCs $1, 2, \ldots, s$. This combining of bit-channels governs the new order of decoding for the combined polar blocks since there is a dependency between two polar blocks due to the combining. At this point, we have got one new bit-channel that became good for all DMCs $1, 2, \ldots, s$ of the combined polar block.

We now see that we can combine many such bit-channels at a time with these two independent polar blocks to achieve universalization. Let us consider bit-channel index set $\mathcal{A}$ and $\mathcal{B}$. Suppose $\mathcal{A}$ has bit-channels that are good for DMCs $1, 2, \ldots, i$ and bad for $i+1, \ldots, s$ whereas $\mathcal{B}$ has bit-channels that are bad for DMCs $1, 2, \ldots, i$ and good for $i+1, \ldots, s$. As already explained, if we consider two polar blocks and combine a bit channel in $\mathcal{A}$ of the first polar block say x with a bit channel in $\mathcal{B}$ of the second polar block then we get two new bit channels, one is bad for all the channels and the other one is good for all channels. In a polar block, a later bit-channel output has the previous bit-channel input as one of its components. So a next valid bit-channel combining could be any bit-channel later than x in $\mathcal{A}$ of the first polar block with any bit-channel later than y in $\mathcal{B}$. Hence a best way to combine bit-channels $\mathcal{A}$ with bit-channels $\mathcal{B}$ is to combine them in order without missing any bit-channels in between. So if $\mathcal{A} = \{x_1, x_2, \ldots, x_{|\mathcal{A}|}\}$, $\mathcal{B} = \{y_1, y_2, \ldots, y_{|\mathcal{B}|}\}$ and G =min$\{|\mathcal{A}|, |\mathcal{B}|\}$, then we do bit-channel combinings x_j with y_j for each $j < G$. This will create a specific order of decoding according to the combining. This is again due to the fact that a later bit-channel output has the previous bit-channel input as one of its components. Now we have G new bit-channels that are good for all channels per updated block length $2N$.

We can consider two such universalized blocks produced in this manner and apply this procedure again. This procedure can be done recursively. So a universalized block obtained after t steps would be of size $2^t N$. We propose to use a hybrid polar block obtained after this universalization procedure in the staircase that satisfies the required condition. We call such a staircase scheme a hybridized staircase scheme. We discuss the average decoding probability of error analysis for the hybridized staircase scheme in the following subsection.

A. Average probability of error analysis for hybridized staircase scheme

We did combining only for bit-channels $\mathcal{H}_X$ of original polar blocks in the first step of the recursive procedure. The bit-channels in $\mathcal{L}_X$ and R are left as is (as shown in Figure 4,) in all the steps of the recursive procedure while producing the hybrid polar block.

Let t be the number of steps in the recursive procedure to produce the hybrid polar block. So we have 2^t original polar blocks in the hybrid polar block. Let us first understand the ensemble average distribution of original polar blocks involved in the hybrid polar block that is used in the staircase scheme. Let us denote the high-entropy bit-channels of the hybrid polar block as $\mathcal{H}_{X_t}$. The bits coded for $\mathcal{H}_{X_t}$ will be uniform iid in the staircase scheme as explained in Section IV. This will assign bits to all bit-channels in $\mathcal{H}_X$ of the original polar blocks involved in the hybrid polar block. The ensemble average distribution of all those bits that go to $\mathcal{H}_X$ in all original polar blocks will be uniform iid for two immediate reasons. One is because combinings form a bijective transform and hence the entropy of bits will be preserved. Second one is because the entropy is maximal for uniform iid distribution and the output of the bijective transformation has an equal number of bits as the input. So the output also has maximal entropy and is uniform iid distributed.

We did not involve the bit-channels in $\mathcal{L}_X$ and R in the combining procedure. We use argmax rule for encoding bit-channels in $\mathcal{L}_X$ as described in Section III. As we use this universalized hybrid polar block in the staircase, for encoding positions in R when the bit-channel lies in a full-height column, we will use randomized rounding rule as explained in Section IV. Notice that all of the original polar blocks in the hybrid polar block are statistically independent of each other according to the average distribution of the random code.

Let $X_1^{1:N}$, $X_2^{1:N}, \ldots, X_{2^t}^{1:N}$ be the codeword components for each of the original polar blocks of a hybrid polar block in the staircase and $Y_1^{1:N}$, $Y_2^{1:N}, \ldots, Y_{2^t}^{1:N}$ be the corresponding received word components when passed through a DMC l selected in S. Let $U_j^{1:N} = X_j^{1:N} G_N$, for $j = 1, 2, \ldots, 2^t$. Now it is clear that the ensemble average distribution of $U_j^{1:N}$ is the same as in the single asymmetric channel case for each j while $\{U_j^{1:N}\}_{j=1}^{2^t}$ are iid distributed. Let us index each hybrid polar block in the staircase as $b = 1, 2, \ldots, Nqk$. Let this actual ensemble average measure on the hybrid polar block b be Q (which is the same for all the hybrid polar blocks in the staircase). Let P be the measure on the hybrid polar block when $X_j^{1:N}$ is iid distributed according to $p(x)$ for each j while $\{X_j^{1:N}\}_{j=1}^{2^t}$ are iid distributed.

If we would like to use the hybrid polar block directly for the universal code without using in staircase, information bits have a uniform iid distribution which are provided to universalized bit-channels. Randomly and independently generated frozen bits with uniform iid distribution are used for other high entropy bit-channels of the hybrid polar block. Boolean functions should be generated for bit-channels in R independently for each of the the original polar blocks involved in the hybrid polar block. They consume exponential storage complexity. Hence the proposed combination of both the schemes above saves these boolean functions. Now it is clear that the ensemble average distribution of hybrid polar block is same as discussed in the above paragraph, here as well.

Lemma 2. *Let $P^j_{X,Y}(x,y)$ be a joint distribution on (X,Y) supported on $\mathcal{X} \times \mathcal{Y}$ for each $j \in \mathcal{J}$. Let $Q(j)$ be the distribution on $\mathcal{J}$. Define $P_{X,Y}(x,y) = \sum_{j\in\mathcal{J}} Q(j)P^j_{X,Y}(x,y)$. Then $Z(X|Y) \geq \sum_{j\in\mathcal{J}} Q(j)Z^j(X|Y)$ where $Z^j(X|Y) = 2\sum_{y\in\mathcal{Y}} \sqrt{P^j_{X,Y}(0,y)P^j_{X,Y}(1,y)}$.*

Proof: Refer to the Appendix.

Lemma 2 is used in the proof of the following proposition.

Proposition 1. *Let (X_1, Y_1) and (X_2, Y_2) be independent random variable pairs which may not be identically distributed. X_1 and X_2 are defined over $\mathcal{X} = \{0, 1\}$, where Y_1 and Y_2 are distributed over alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$. Let $U_1 = X_1 + X_2$ and $U_2 = X_2$. Then*

1. $Z(U_1|Y_1Y_2) \geq \max\{Z(X_1|Y_1), Z(X_2|Y_2)\}$.

2. $Z(U_2|U_1Y_1Y_2) = Z(X_1|Y_1)Z(X_2|Y_2)$.

Proof: Refer to the Appendix.

Lemma 3. *Let $P_1(x_1)$ and $Q_1(x_1)$ be distributions on random variables X_1 and assume the total variation distance between P_1 and Q_1 is ϵ_1. Let P_2 and Q_2 be distributions on random variable X_2 and assume the total variation distance between $P_2(x_2)$ and $Q_2(x_2)$ is ϵ_2. Now $P(x_1, x_2) = P_1(x_1)P_2(x_2)$ and $Q(x_1, x_2) = Q_1(x_1)Q_2(x_2)$ become two joint distributions on random variables X_1 and X_2 such that both the random variables are mutually independent. Then the total variation distance between the distributions P and Q is at most $\epsilon_1 + \epsilon_2$.*

Proof: Refer to the Appendix.

Lemma 4. *Let the (X, Y) random variable pair have two measures defined as $Q(x, y) =$*

$Q(x)p(y|x)$ and $P(x,y) = P(x)p(y|x)$, respectively. So the conditional distributions $Q(y|x)$ and $P(y|x)$ are both equal to $p(y|x)$. Now the total variation between the joint distributions $||Q_{X,Y} - P_{X,Y}||$ becomes $||Q_X - P_X||$.

Proof: Refer to the Appendix.

The analysis of asymptotic probability of error is as follows:

- Let us denote the low entropy bit-channels of the hybrid polar block as $\mathcal{L}_{X_t}$. Let $\mathcal{E}_b = \cup_{i \in I_l \cup \mathcal{L}_{X_t}} \mathcal{E}_{ib}$ be the error event corresponding to a hybrid polar block b in the staircase. Here I_l is the updated good bit-channel set of a hybrid polar block and $\mathcal{E}_{ib}$ is the error event of bit-channel i for hybrid polar block b. Notice that $\mathcal{E}_b$ is not the actual decoding error event for the hybrid polar block in the staircase since bit corruption in bit-channels $\mathcal{H}_{X_t} - \cup_{i \in I_l \cup \mathcal{L}_{X_t}}$ may occur due to errors in other hybrid polar blocks while RS decoding. Now the average probability of error $Q(\mathcal{E}_b)$ can be bounded as the sum of two entities as follows.
$$Q(\mathcal{E}_b) \leq P(\mathcal{E}_b) + ||P - Q||$$
- The Bhattarcharya parameters of the updated good bit-channels I_l of hybrid polar block in measure P is $O(2^{-N^{\beta}})$ where $\beta < 0.5$. This is deduced by the recursive application of Proposition 1. So the sum of all such Bhattarcharya parameters is at most $O(2^t 2^{-N^{\beta}})$. So $P(\mathcal{E}_b)$ is upper bounded by $O(2^t 2^{-N^{\beta}})$ using the union bound. This accounts for one of the sum components of the upper bound on $Q(\mathcal{E}_b)$.
- Now, the total variation distance satisfies
$$\begin{aligned} ||P - Q|| &\overset{(a)}{\leq} \sum_{j=1}^{2^t} ||P_{X_j^{1:N}, Y_j^{1:N}} - Q_{X_j^{1:N}, Y_j^{1:N}}|| \\ &\overset{(b)}{=} \sum_{j=1}^{2^t} ||P_{X_j^{1:N}} - Q_{X_j^{1:N}}|| \end{aligned}$$
The identity (a) is due to the fact that $\{X_j^{1:N}\}_{j=1}^{2^t}$ are iid distributed in both the measures P,Q and also due to Lemma 3. The identity (b) is true by the application of Lemma 4 coupled with the fact that conditional measure of $Y_j^{1:N}$ given $X_j^{1:N}$ is same in both the P and Q measures that is induced by the selected DMC in S. Now the total variation distance $||P_{X_j^{1:N}} - Q_{X_j^{1:N}}||$ is $O(2^{-N^{\beta}})$ as we discussed in Section III for the single asymmetric channel. Overall $||P - Q||$ is again upper bounded by $O(2^t 2^{-N^{\beta}})$.

- Hence $Q(\mathcal{E}_b)$ is upper bounded by $O(2^{t+1}2^{-N^\beta})$ for any DMC l in S.
- As mentioned in the Theorem 1, the overall error event becomes $\mathcal{E} = \cup_{b\in\{1:Nqk\}}\cup_{i\in\mathcal{L}_{X_t}\cup I_l}\mathcal{E}_{ib}$.
- The overall average error probability is upper bounded by $\sum_{b\in\{1:Nqk\}} Q(\mathcal{E}_b)$, as we did in Theorem 1 using union bound. Hence the overall average error probability of the hybridized staircase scheme will become $O(Nqk2^{t+1}2^{-N^\beta})$.

If one would like to do the probability of error analysis for the universal code by just using the hybrid polar block without the staircase scheme, we can mimic the above steps as we had the same ensemble average distribution on hybrid polar block. In Section VI, we will illustrate the scheme in more detail with the universal block produced after the first step of the recursive combining procedure in the case of $S = \{1, 2\}$.

Remark: We would also like to mention that the chaining construction for universalization provided in [6] can be extended for asymmetric case (where the component codewords of the original polar blocks involved in the construction become dependent). The same set of boolean functions could also be used for all original polar blocks involved in universalization for not-completely-polarized bit-channels R since the statistical independence is absent between original blocks. The probability of error in the asymmetric case would be $O(2w2^{-N^\beta})$ where w is the number of blocks involved in the construction.

B. Algorithm to produce hybrid polar block for the staircase scheme

Now we provide a method to produce the hybrid polar block that achieves the required condition with a blocklength that is at most 2^{s-1} times the original polar block length.

If $|I_1 \cap I_2 \cap \ldots \cap I_{s-1}| \geq 2|R|$, then we consider two polar blocks. We combine "bit channels that are good for $1, 2, \ldots, s-1$ and bad for s" of one block with "bit channels that are good for DMC s and bad for DMCs $1, 2, \ldots, s-1$" of the other block. Hence we get "min$\{|I_1 \cap I_2 \cap \ldots \cap I_{s-1}| - |I_1 \cap I_2 \cap \ldots \cap I_s|, |I_s| - |I_1 \cap I_2 \cap \ldots \cap I_s|\}$" number of good bit-channels for all DMCs from $1, 2, \ldots, s$. The improvement per block length N will be half of the above quantity which is at least $\frac{2|R|-|I_1\cap I_2\cap\ldots\cap I_s|}{2}$. So the updated $|I_1 \cap I_2 \cap \ldots \cap I_s|$ per block length N would be $|R| + |I_1 \cap I_2 \cap \ldots \cap I_s|/2$, which is greater than or equal to $|R|$. Hence our condition is met for the block to be used in the staircase scheme.

If $|I_1 \cap I_2 \ldots \cap I_{s-1}| < 2|R|$, we do the same kind of combining step explained in the above paragraph at this step and get the required condition $|I_1 \cap I_2 \cap \ldots \cap I_{s-1}| > 2|R|$ to execute the step in the previous paragraph. This is a recursive procedure and the required condition will be guaranteed at some point because as the recursive procedure hits $s-1$ steps, we will have $\min\{|I_1|, |I_2|, \ldots |I_s|\} > 2^s|R|$ for large enough block length. So we have at most $s-1$ steps in the recursive procedure and so the maximum block length will be at most $2^{s-1}N$.

Algorithm 1 and 2 together will provide the high-level representation of the recursive procedure just mentioned for producing the required hybrid polar block to be used in the staircase. By calling Algorithm 1, we get the required hybrid polar block of size at most $2^{s-1}N$.

Algorithm 1 : GetHybridizedPolar Block

```
procedure GETHYBRIDIZEDPOLARBLOCK
    Initialize $currentType$ as per original polar block.        ▷ variable to update polar block
    Set $S = \{1, 2, ..., s\}$
    if ($| \cap_{i \in S} I_i| < |R|$)        ▷ $| \cap_{i \in S} I_i|$ denotes size per block length $N$ of $currentType$
        recursivelyUpdateBlock(0,$S$);
    end if
    return
```

Algorithm 2 : recursivelyUpdateBlock($recursiveCount$, $channelsInvolved$)

```
procedure RECURSIVELYUPDATEBLOCK
    Set $S = \text{argmax}_{S \subset channelsInvolved : |S| = |channelsInvolved|-1} | \cap_{i \in S} I_i|$ ;
    if ($| \cap_{i \in S} I_i| < 2^{recursiveCount+1}|R|$)
        recursivelyUpdateBlock($recursiveCount + 1$,$S$);
    end if
    Consider two polar blocks of $currentType$
    Combine bit channels that are "good for $S$ and bad for $channelsInvolved - S$ " of one
      block with bit channels that are "bad for $S$ and good for $channelsInvolved - S$ " of the
      other block to produce the updated polar block;
    Update the global variable $CurrentType$ as per the updated polar block that has the size
      $| \cap_{i \in channelsInvolved} I_i| \geq 2^{recursiveCount}|R|$
    return;
```

The global variable $currentType$ used in Algorithm 1 and 2 denotes the updated polar block. The size $|\cap_{i\in S} I_i|$ in Algorithm 1 and 2 gives the cardinality per blocklenth N according to the updated polar block represented by $currentType$. Note that the recursive flow in Algorithm 2 is different from the recursive flow of the original scheme presented in [5]. In our algorithm, the polar block used in the kth step is the polar block returned from the $k+1$th recursive call. That is why, the given algorithm has recursive call first and then the combining step for the universalization. But in the original scheme [5], the polar block used in the kth step is the block produced after $k-1$th step. So the original scheme should have the combining step first followed by the recursive call, in contrast to our proposed algorithm. Also note that the recursion depth in the presented algorithm is at most $s-1$ as we have the base condition min$\{|I_1|, |I_2|, ...|I_s|\} > 2^s|R|$ satisfied for the original polar block.

The combining in step 7 of Algorithm 2 would give at least $|\cap_{i\in S}I_i| - |\cap_{i\in channelsInvolved}I_i|$ new bit-channels that are good for all channels in $channelsInvolved$ per block length $2N$. So the updated $|\cap_{i\in channelsInvolved}I_i|$ would be at least the sum of the present values of $|\cap_{i\in channelsInvolved}I_i|$ and $\frac{|\cap_{i\in S}I_i| - |\cap_{i\in channelsInvolved}I_i|}{2}$. So it would be at least $\frac{|\cap_{i\in S}I_i| + |\cap_{i\in channelsInvolved}I_i|}{2}$ which would be at least $2^{recursiveCount}|R|$. This should be the required condition before returning to the previous recursive call.

We may improve the Algorithm 2 to have lesser recursive calls. Nevertheless the mentioned recursive procedure guarantees a block length which will be at most $2^{s-1}N$. Hence the overall block length of the hybridized staircase scheme would be $O(2^{s-1}qkN^2)$.

VI. Universal scheme based on combining different bit-channels

In Section V, we analysed the probability of error considering a block produced by this universal scheme in the context of hybridized staircase scheme. We now describe in detail the code construction, including encoding and decoding methods in the case of $S = \{1, 2\}$ for the block produced after the first step in the recursive procedure.

A. Combining two polar blocks for universalizing with respect to both the DMCs in S

Let $G = \min\{|I_1 \cap F_2|, |I_2 \cap F_1|\}$. Consider the sets $\mathcal{A}$ and $\mathcal{B}$ to be the first G indices in $I_1 \cap F_2$ and $I_2 \cap F_1$ respectively. Let $\mathcal{A} = \{x_1, x_2, \ldots, x_G\}$ where $x_1 < x_2 < \ldots < x_G$ and $\mathcal{B} = \{y_1, y_2, \ldots, y_G\}$ where $y_1 < y_2 < \ldots < y_G$. Consider two independent polar blocks, say block 1 and block 2, of length N. Let $X^{1:2N}$ be iid distributed according to compound capacity-achieving $p(x)$ and

$$U^{1:N} = X^{1:N}G_N, \quad V^{1:N} = X^{N+1:2N}G_N \tag{1}$$

Let us associate $X^{1:N}$ and $X^{N+1:2N}$ to block 1 and block 2, respectively. Now we combine bit-channel x_j of block 1 with y_j of block 2, which produces two new bit-channels for each $j \in \{1 : G\}$ with inputs $U'_{x_j} = U_{x_j} + V_{y_j}$ and $V'_{y_j} = V_{y_j}$. For the remaining bit-channels which are neither x_j for block 1 nor y_j for block 2, the same bit-channels are used. So we have $U'_i = U_i$ and $V'_i = V_i$ for these bit-channels. This is shown in Figure 4. Let $\{(X_i, Y_i)\}_{i=1}^{2N}$ be iid distributed according to $p(x)p_l(y|x)$ for $l \in S$.

Lemma 5. *For each $j \in \{1 : G\}$, for any $\beta' < \beta < 0.5$, for sufficiently large N*

1. $Z(U'_{x_j}|U^{1:x_j-1}V^{1:y_j-1}Y^{1:2N}) \geq 1 - 2^{-N^{\beta'}}$.

(U'_{x_i} is almost random given $U^{1:x_j-1}V^{1:y_j-1}Y^{1:2N}$ for both DMCs $l = 1, 2$)

2. $Z(V'_{y_i}|U^{1:x_i-1}V^{1:y_i-1}U'_{x_i}Y^{1:2N}) \leq 2^{-N^{\beta'}}$.

(V'_{y_i} is almost deterministic given $U^{1:x_j-1}U'_{x_j}V^{1:N+y_j-1}Y^{1:2N}$ for both DMCs $l = 1, 2$)

Proof: Refer to the Appendix.

It follows from Lemma 5 that the bit-channel combinings mentioned above lead to $2|I_1 \cap I_2| + G$ new good bit-channels with respect to block length $(2N)$.

The proposed random code construction makes $(U'^{1:N}, V'^{1:N})$ such that the ensemble average distribution of $(U'^{1:N}, V'^{1:N})$ will be $O(2^{-N^{\beta'}})$ close to the distribution when the word $X^{1:2N}$ is iid distributed according to $p(x)$. Now we describe the code construction.

B. Code construction

We first generate random functions $f_1 : \mathcal{H}_X - (I_1 \cap I_2) \rightarrow \{0, 1\}$ and $f_2 : \mathcal{H}_X - ((I_1 \cap I_2) \cup \mathcal{B}) \rightarrow \{0, 1\}$ where each $f_i(j)$, $j \in F$ and $i \in \{1, 2\}$ is chosen independently and uniformly. These frozen bits are shared between encoder and decoder.

For both the block $b = 1$ and $b = 2$, we generate independent random boolean functions $\lambda_i^b : \{0, 1\}^{i-1} \rightarrow \{0, 1\}$ for each $i \in R$ by using the following probability rule:

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

independently for each $u^{1:i-1}$. Let the set of random functions be denoted by λ_R^b. These functions are shared between encoder and decoder. Now we describe the encoding and decoding algorithms.

Encoding

Input: uniformly distributed message M of $2|I_1 \cap I_2| + G$ bits

Output: codeword $X^{1:2N}$

for $i = 1 : N$, encode U_i as follows.

1. Partition M into M_1 and M_2 such that M_1 takes the first $|I_1 \cap I_2|$ bits of M and M_2 takes the last $|I_1 \cap I_2| + G$ bits of M. Set $U'^{I_1 \cap I_2} = M_1$ and $V'^{(I_1 \cap I_2) \cup \mathcal{B}} = M_2$.

2. Set $U'_i = f_1(i)$ for $i \in \mathcal{H}_X - ((I_1 \cap I_2))$ and $V'_i = f_2(i)$ for $i \in \mathcal{H}_X - ((I_1 \cap I_2) \cup \mathcal{B})$ in block 1 and block 2 respectively.

3. Now $U_{x_j} = U'_{x_j} + V'_{y_j}$ and $V_{y_j} = V'_{y_j}$ for $j \in \{1 : G\}$. $U_i = U'_i$ and $V_i = V'_i$ for other $i \in \mathcal{H}_X$.

4. For $i \in \mathcal{L}_X$ for block 1 and block 2, we use the following argmax rule for encoding these bit-channels.

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

$$V_i = \text{argmax}_{x \in \{0,1\}} P_{V_i|V^{1:i-1}}(x|V^{1:i-1})$$

5. We assign $U_i = \lambda_i^1(U^{1:i-1})$ and $V_i = \lambda_i^2(V^{1:i-1})$ for encoding these indices. For $i \notin \mathcal{H}_X$, take $U'_i = U_i$ and $V'_i = V_i$. Transmit $X^{1:N} = U^{1:N} G_N$ and $X^{N+1:2N} = V^{1:N} G_N$.

Decoding

Input: received vector $Y^{1:2N}$

Output: message estimate $\hat{M}$ of $1 : 2|I_1 \cap I_2| + G$ bits

```
1. Set $j = 1$, Set $x_0 = 0$ and $y_0 = 0$.
2. for $i = x_{j-1} + 1 : x_j - 1$ of block 1
    If $i \in \mathcal{H}_X - (I_1 \cap I_2)$ the set
    $\hat{U}'_i = \hat{U}_i = f_1(i)$.
    If $i \in (I_1 \cap I_2) \cup \mathcal{L}_X$ then set
      $\hat{U}'_i = \hat{U}_i = \mathbb{1}\{P_{U_i|U^{1:i-1},Y^{1:N}}(1|\hat{U}^{1:i-1}, Y^{1:N}) \geq P_{U_i|U^{1:i-1},Y^{1:N}}(0|\hat{U}^{1:i-1}, Y^{1:N})\}$.
    If $i \in (\mathcal{H}_X \cup \mathcal{L}_X)^c$ then set
    $\hat{U}'_i = \hat{U}_i = \lambda_i^1(\hat{U}^{1:N-1})$.
    end
    for $i = y_{j-1} + 1 : y_j - 1$ of block 2
```

```
        If $i \in \mathcal{H}_X - (I_1 \cap I_2)$ then set
        $\hat{V}_i' = \hat{V}_i = f_2(i)$.
        If $i \in (I_1 \cap I_2) \cup \mathcal{L}_X$ then set
        $\hat{V}_i' = \hat{V}_i = \mathbb{1}\{P_{U_i|U^{1:i-1},Y^{1:N}}(1|\hat{V}^{1:i-1}, Y^{N+1:2N}) \geq P_{U_i|U^{1:i-1},Y^{1:N}}(0|\hat{V}^{1:i-1}, Y^{N+1:2N})\}$.
        If $i \in (\mathcal{H}_X \cup \mathcal{L}_X)^c$ then set
        $\hat{V}_i' = \hat{V}_i = \lambda_i^2(\hat{V}^{1:N-1})$.
    end
3. Set
    $\hat{U}_{x_j}' = f_1(x_j)$.
    $\hat{V}_{y_j}' = \mathbb{1}\{P_{V_{y_j}'|U^{1:x_j-1}U_{x_j}'V^{1:y_j-1}Y^{1:2N}}(1|\hat{U}^{1:x_j-1}U_{x_j}'\hat{V}^{1:y_j-1}Y^{1:2N}) \geq P_{V_{y_j}'|U^{1:x_j-1}U_{x_j}'V^{1:y_j-1}Y^{1:2N}}(0|$
                                        $\hat{U}^{1:x_j-1}\hat{U}_{x_j}'\hat{V}^{1:y_j-1}Y^{1:2N})\}$.
    $\hat{U}_{x_j} = \hat{U}_{x_j}' + \hat{V}_{y_j}'$ and $\hat{V}_{y_j} = \hat{V}_{y_j}'$.
4. Repeat step 2 and 3 for $j = \{2, 3, \ldots, G\}$.
5. Set $\hat{M}_1 = \hat{U}'^{I_1 \cap I_2}$ and $\hat{M}_2 = \hat{V}'^{(I_1 \cap I_2) \cup \mathcal{B}}$. Combine $\hat{M}_1$, $\hat{M}_2$ to get $\hat{M}$.
```

Step 2 in the above decoding algorithm is to decode bit-channels, which are not involved in the combining process, where as the step 3 is to decode the new bit-channels, which are produced after combining.

Theorem 2. *Let $P_{e,l}(\lambda_R^1, \lambda_R^2, f_1, f_2)$ denotes the decoding probability of error when DMC l is selected in S for a given code in the above random code construction. For sufficiently large block length N, the average decoding probability of error $\mathbb{E}[P_{e,l}(\lambda_R^1, \lambda_R^2, f_1, f_2)] = O(2^{-N^{\beta'}})$ for any $\beta' < \beta < 0.5$ and for each $l \in S$.*

Proof: Refer to the Appendix.

VII. 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 under an assumption on the cardinality of the intersection of good-bit channels. When the assumption does not hold, we proposed a hybridized staircase scheme with a block length that is at most 2^{s-1} times the original polar block length that satisfies condition on the cardinality of the intersection of good-bit channels. The scheme requires a large block length, leaving open the problem of designing codes with short block length. Another open problem could be the construction of a stronger universal polar code with reduced storage complexity that 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.

VIII. 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] A. El Gamal and Y. H. Kim, *Network Information Theory*, Cambridge, UK: Cambridge University Press, 2011.

[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:

1.

We prove part 1 of Theorem 1 in the following steps.

Step 1:

Consider any polar block in the extended staircase which lies completely in the full-height column regime. This polar block corresponds to N distinct full-height columns. To get the distribution on encoded codeword $U^{1:N}$ for such a polar 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 each bit-channel i in the block.

If $i \in \mathcal{L}_X$, they 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 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

$$\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N}) = 0.5.$$

Now we discuss the case $i \in I'$ in detail. Let the parity of information bits corresponding to that column be X. Let Y be the already encoded bit in the block corresponding to the bit-channel $g(i)$ in that column. Here we assigned U_i as $X \oplus Y$. Here the distribution of X will be Bernoulli(0.5). Let the distribution of Y be Bernoulli(p) for some p. Note that both the random variables X and Y are independent.

Observe that

$$\begin{aligned}\mathbb{P}(U_i = x|Y = y, W^{1:N}) &= \mathbb{P}(X + Y = x|Y = y, W^{1:N})\\ &= \mathbb{P}(X = x + y|Y = y, W^{1:N})\\ &\overset{(a)}{=} \mathbb{P}(X = x + y)\\ &= 0.5.\end{aligned}$$

Identity (a) is true since X and Y are independent. We established the independence of U_i and Y. Now we establish that U_i is independent of all the encoded bits of previous columns and the frozen vector $W^{1:N}$. This will imply that U_i is be independent of $U^{1:i-1}$ of that block. Let $\overline{P}$

be the random vector denoting the encoded bits of the previous columns. Now the conditional probability

$$\begin{aligned}
\mathbb{P}(U_i = u_i|\overline{P}, W^{1:N}) &= \sum_{y\in\{0,1\}} \mathbb{P}(U_i = u_i, Y = y|\overline{P}, W^{1:N}) \\
&= \sum_{y\in\{0,1\}} \mathbb{P}(Y = y|\overline{P}, W^{1:N})\mathbb{P}(U_i = u_i|Y = y, \overline{P}, W^{1:N}) \\
&\overset{(a)}{=} \sum_{y\in\{0,1\}} \mathbb{P}(Y = y|\overline{P}, W^{1:N})\mathbb{P}(U_i = u_i|Y = y) \\
&\overset{(b)}{=} \sum_{y\in\{0,1\}} \mathbb{P}(Y = y|\overline{P}, W^{1:N})\mathbb{P}(U_i = u_i) \\
&= \mathbb{P}(U_i = u_i) \\
&= 0.5.
\end{aligned} \tag{2}$$

Identity (a) is true since U_i is independent of encoded bits of previous columns, given the random variable Y. Identity (b) by independence of U_i with respect to Y. Hence the distribution of $U^{1:N}$ for a block encoded in full-height column regime becomes

$$\begin{aligned}
\mathbb{P}(U^{1:N} = u^{1:N}|W^{1:N}) &= \Pi_{i\in[N]}\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N}) \\
&= 2^{-|\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}).
\end{aligned} \tag{3}$$

This implies that

$$\mathbb{E}_{W^{1:N}}[\mathbb{P}(U^{1:N} = u^{1:N}|W^{1:N})] = 2^{-|\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}). \tag{4}$$

Step 2:

Consider a polar block which lies partly in the non-full-height column regime on the right side. For U_i in a full-height column, the conditional probability rule is already derived in step 1. We now derive the conditional probablility for U_i in a non-full-height column.

If $i \in \mathcal{H}_X$, then by encoding rule

$$\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N}) = 0.5.$$

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 encoding rule

$$\mathbb{P}(U_i = u_i|U^{1:i-1} = u^{1:i-1}, W^{1:N}) = P_{U_i|U^{1:i-1}W^{1:N}}(u_i|u^{1:i-1}).$$

It implies that

$$\begin{aligned}\mathbb{P}(U^{1:N}=u^{1:N}|W^{1:N}) &= \Pi_{i\in[N]}\mathbb{P}(U_i=u_i|U^{1:i-1}=u^{1:i-1},W^{1:N})\\ &= 2^{-|\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}).\end{aligned} \tag{5}$$

The ensemble average distribution becomes

$$\mathbb{E}_{W^{1:N}}[\mathbb{P}(U^{1:N}=u^{1:N}|W^{1:N})] = 2^{-|\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}). \tag{6}$$

Step 3:

Consider a polar block which lies partly in the non-full-height column regime on the left side. Let H be the indices of the block that lies in non-full-height columns. Those bits are encoded as W_i corresponding to every index $i \in H$.

Now,

$$\begin{aligned}\mathbb{E}_{W^{1:N}}[\Pi_{i\in H}\mathbb{1}(u_i=W_i)] &= \mathbb{E}_{W^{1:N}}[\mathbb{1}(\cap_{i\in H}(u_i=W_i))]\\ &= \mathbb{E}_{W^{1:N}}[\sum_{u_i\in\{0,1\}:i\in H^c}\mathbb{1}(\cap_{i\in[N]}(u_i=W_i))]\\ &= \sum_{u_i\in\{0,1\}:i\in H^c}\mathbb{E}_{W^{1:N}}[\mathbb{1}(\cap_{i\in[N]}(u_i=W_i))].\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}_{W^{1:N}}[\Pi_{i\in H}\mathbb{1}(u_i=W_i)] &\overset{(a)}{=} \sum_{u_i\in\{0,1\}:i\in H^c} 2^{-|\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R}\\ &\qquad P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}).\\ &= 2^{-|H\cap\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X\cap H}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R\cap H}\\ &\qquad P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}).\end{aligned} \tag{7}$$

Identity (a) is true because we have

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

from our random code construction. Now we compute the distribution of $U^{1:N}$

$$\begin{aligned}\mathbb{P}(U^{1:N}=u^{1:N}|W^{1:N}) &= \Pi_{i\in[N]}\mathbb{P}(U_i=u_i|U^{1:i-1}=u^{1:i-1},W^{1:N})\\ &\overset{(a)}{=} \Pi_{i\in H}\mathbb{1}(u_i=W_i)2^{-|\mathcal{H}_X-H|}\Pi_{i\in\mathcal{L}_X-H}\delta_i(u_i|u^{1:i-1})\\ &\qquad \Pi_{i\in R-H}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}).\end{aligned} \tag{8}$$

Identity (a) is true since the conditional distribution of U_i given $U^{1:i-1}$ in a non-full-height column is derived in step 1.

This implies

$$\begin{aligned}
\mathbb{E}_{W^{1:N}}[\mathbb{P}(U^{1:N}=u^{1:N}|W^{1:N})] &= \mathbb{E}_{W^{1:N}}[\Pi_{i\in H}\mathbb{1}(u_i=W_i)2^{-|\mathcal{H}_X-H|}\Pi_{i\in\mathcal{L}_X-H} \\
&\quad \delta_i(u_i|u^{1:i-1})\Pi_{i\in R-H}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1})] \\
&= \mathbb{E}_{W^{1:N}}[\Pi_{i\in H}\mathbb{1}(u_i=W_i)]2^{-|\mathcal{H}_X-H|}\Pi_{i\in\mathcal{L}_X-H} \qquad (9) \\
&\quad \delta_i(u_i|u^{1:i-1})\Pi_{i\in R-H}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}) \\
&\overset{(a)}{=} 2^{-|\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}).
\end{aligned}$$

Identity (a) follows from equation (7). We now have the required ensemble average distribution of $U^{1:N}$ for these blocks as well. This concludes the proof of part 1.

2.

Let $\mathcal{E}$ be the error event and l be the DMC selected in S. The error occurs if and only if there is a decoding error while decoding some bit-channel in $\mathcal{L}_X\cup I_l$ of any polar block in the staircase. Let us index each polar block in the staircase as $b=1,2,\ldots,Nqk$. Let $\mathcal{E}_g$ be the error event with a genie-aided decoder which has the accurate values of the past $U^{1:i-1}$ when decoding any bit-channel $i\in\mathcal{L}_X\cup I_l$ for all polar blocks. Let $\mathcal{E}_{ib}$ be the bit-channel error for the bit-channel i corresponding to the block b. Note that for a bit-channel i that lies in the full-height column of b,

$$\begin{aligned}
\mathcal{E}_{ib} = \{&(U^{1:N},Y^{1:N}) \text{ of block } b\ : \\
&P_{U_i|U^{1:i-1},Y^{1:N}}(U_i|U^{1:i-1},Y^{1:N}) \geq P_{U_i|U^{1:i-1},Y^{1:N}}(U_i|U^{1:i-1},Y^{1:N})\}.
\end{aligned}$$

If the bit-channel i that lies in the non-full-height column of polar block b, then $\mathcal{E}_{ib}$ will be the null event. Clearly we can deduce that $\mathcal{E}_g=\cup_{b\in\{1:Nqk\}}\cup_{i\in\{1:N\}}\mathcal{E}_{ib}$. Note that error event $\mathcal{E}$ will imply at least one of the error events $\{\mathcal{E}_{ib}:b\in\{1:Nqk\}$ and $i\in\{1:N\}\}$. So we should have

$$\mathcal{E}\subset\mathcal{E}_g.$$

One the other hand, it is obvious that

$$\mathcal{E}_g\subset\mathcal{E}.$$

Hence the error event $\mathcal{E}$ will be as below:

$$\mathcal{E}=\cup_{b\in\{1:Nqk\}}\cup_{i\in\mathcal{L}_X\cup I_l}\mathcal{E}_{ib}.$$

The application of the union-bound followed by taking the expectation of the random ensemble of codes gives the following upper-bound for ensemble average probability of error:

$$\mathbb{E}_{W^{1:N}}[\mathbb{P}(\mathcal{E}|W^{1:N})] \leq \sum_{b\in\{1:Nqk\}} \mathbb{E}_{W^{1:N}}[\mathbb{P}(\cup_{i\in\mathcal{L}_X\cup I_l}\mathcal{E}_{ib}|W^{1:N})].$$

Now the ensemble average distribution is $O(2^{-N^{\beta}})$ close to the distribution when $X^{1:N}$ is iid according to $p(x)$ in total variation distance for each of these blocks in all the staircases.

Therefore the average probability of error $\mathbb{E}_{W^{1:N}}[\mathbb{P}(\cup_{i\in\mathcal{L}_X\cup I_l}\mathcal{E}_{ib}|W^{1:N})]$ becomes $O(2^{-N^{\beta'}})$ for $\beta' < 0.5$ as derived in [7]. Hence the overall ensemble average probability of error will be $O(Nqk2^{-N^{\beta'}})$. This concludes the proof of part 2.

3.

Encoding Complexity: Encoding complexity consists of two factors: encoding the polar block and encoding the RS codeword. Encoding the polar block takes $O(N\log_2(N))$ real operations. Hence the number of operations per bit would be $O(\log_2(N))$ real operations. Encoding a RS codeword can done in $O(N\log_2(N))$ operations over GF(2^q). Addition and multiplication over this field take q and $q^{\log_2(3)}$ binary operations, respectively. Hence there can be maximum of $q^{\log_2(3)}O(N\log_2(N))$ binary operations. One RS codeword contains $L'q$ bits. Hence the encoding takes $O((log_2N)q^{log_2(3)-1})$ binary operations per bit.

Decoding Complexity:Decoding complexity consists of two factors: decoding the polar block and decoding the RS codeword. Decoding the polar blocks takes $O(N\log_2 N)$ real operations. Hence the number of operations per bit would be $O(\log_2(N))$ real operations. Decoding a RS codeword can be done in $O(N(\log_2(N))^2)$ operations over GF(2^q). Addition and multiplication over this field take q and $q^{\log_2(3)}$ binary operations, respectively. One codeword contains $L'q$ bits. Hence there will be maximum of $q^{\log_2(3)}O(N(\log_2(N))^2)$ binary operations. Hence the decoding takes $O(N(\log_2(N))^2)q^{log_2(3)-1})$ binary operations per bit. This concludes the proof of part 3. □

Proof of Lemma 2:

$$\begin{aligned}Z(X|Y) &= 2\sum_{y\in\mathcal{Y}}\sqrt{P_{X,Y}(0,y)P_{X,Y}(1,y)}\\ &= -1+\sum_{y\in\mathcal{Y}}\big[\sum_{x\in\mathcal{X}}\sqrt{P_{X,Y}(x,y)}\big]^2\end{aligned}$$

$$\begin{aligned}&\overset{(a)}{\geq} -1+\sum_{y\in\mathcal{Y}}\sum_{j\in\mathcal{J}}Q(j)\big[\sum_{x\in\mathcal{X}}\sqrt{P^{j}_{X,Y}(x,y)}\big]^2\\ &=\sum_{j\in\mathcal{J}}Q(j)(-1+\sum_{y\in\mathcal{Y}}\big[\sum_{x\in\mathcal{X}}\sqrt{P^{j}_{X,Y}(x,y)}\big]^2)\\ &=\sum_{j\in\mathcal{J}}Q(j)Z^j(X|Y).\end{aligned} \tag{10}$$

Identity (a) is true by the following Minkowsky's inequality, when $r<1$ and a_{jk} is non-negative:

$$\sum_{k\in\mathcal{K}}(\sum_{j\in\mathcal{J}}Q(j){a_{jk}}^{\frac{1}{r}})^r \geq \big[\sum_{j\in\mathcal{J}}Q(j)(\sum_{k\in\mathcal{K}}a_{jk})^{\frac{1}{r}}\big]^r.$$

Here $r=0.5$ and $a_{jk}=\sqrt{P^{j}_{X,Y}(x,y)}$. □

Proof of Proposition 1:

The conditional distribution $P_{U_1|Y_1Y_2}(u_1|y_1y_2)$ will be as follows:

$$\begin{aligned}P_{U_1|Y_1Y_2}(u_1|y_1y_2)&=\sum_{u_2\in\mathcal{X}}P_{U_1,U_2|Y_1Y_2}(u_1,u_2|y_1y_2)\\ &=\sum_{u_2\in\mathcal{X}}P_{X_1X_2|Y_1Y_2}(u_1+u_2,u_2|y_1y_2)\\ &=\sum_{u_2\in\mathcal{X}}P_{X_1|Y_1}(u_1+u_2|y_1)P_{X_2|Y_2}(u_2|y_2).\end{aligned}$$

The conditional distribution $P_{U_2|Y_1Y_2U_1}(u_2|y_1y_2u_1)$ will be as follows:

$$\begin{aligned}P_{U_2|Y_1Y_2U_1}(u_2|y_1y_2u_1)&=\frac{P_{U_1,U_2|Y_1Y_2}(u_1,u_2|y_1y_2)}{P_{U_1|Y_1Y_2}(u_1|y_1y_2)}\\ &=\frac{P_{X_1|Y_1}(u_1+u_2|y_1)P_{X_2|Y_2}(u_2|y_2)}{\sum_{u_2\in\mathcal{X}}P_{X_1|Y_1}(u_1+u_2|y_1)P_{X_2|Y_2}(u_2|y_2)}.\end{aligned}$$

The joint distribution $P_{U_1Y_1Y_2}(u_1,y_1,y_2)$ will be as follows:

$$\begin{aligned}P_{U_1Y_1Y_2}(u_1,y_1,y_2)&=\sum_{u_2\in\mathcal{X}}P_{U_1U_2Y_1Y_2}(u_1,u_2,y_1,y_2)\\ &=P_{U_1U_2Y_1Y_2}(u_1,0,y_1,y_2)+P_{U_1U_2Y_1Y_2}(u_1,1,y_1,y_2)\\ &=P_{U_2}(0)P_{U_1Y_1Y_2|U_2}(u_1,y_1,y_2|0)+P_{U_2}(1)P_{U_1Y_1Y_2|U_2}(u_1,y_1,y_2|1).\end{aligned}$$

Let

$$P^1_{U_1Y_1Y_2}(u_1,y_1,y_2)=P_{U_1Y_1Y_2|U_2}(u_1,y_1,y_2|0),$$

$$P^2_{U_1Y_1Y_2}(u_1,y_1,y_2)=P_{U_1Y_1Y_2|U_2}(u_1,y_1,y_2|1)$$

be the two joint distributions on random variable triplet (U_1, Y_1, Y_2). We now derive the Bhattacharya parameter corresponding to the distribution $P^1_{U_1Y_1Y_2}(u_1, y_1, y_2)$.

$$\begin{aligned}
Z^1(U_1|Y_1Y_2) &= 2\sum_{y_1y_2} P^1_{Y_1Y_2}(y_1y_2)\sqrt{P^1_{U_1|Y_1Y_2}(0|y_1y_2)P^1_{U_1|Y_1Y_2}(1|y_1y_2)} \\
&= 2\sum_{y_1y_2} P_{Y_1Y_2|U_2}(y_1y_2|0)\sqrt{P_{U_1|Y_1Y_2U_2}(0|y_1y_20)P_{U_1|Y_1Y_2U_2}(1|y_1y_20)} \\
&\overset{(a)}{=} 2\sum_{y_1y_2} P_{Y_1}(y_1)P_{Y_2|U_2}(y_2|0)\sqrt{P_{X_1|Y_1Y_2U_2}(0|y_1y_20)P_{X_1|Y_1Y_2U_2}(1|y_1y_20)} \\
&\overset{(b)}{=} 2\sum_{y_1y_2} P_{Y_1}(y_1)P_{Y_2|U_2}(y_2|0)\sqrt{P_{X_1|Y_1}(0|y_1)P_{X_1|Y_1}(1|y_1)} \\
&= 2\sum_{y_1} P_{Y_1}(y_1)\sqrt{P_{X_1|Y_1}(0|y_1)P_{X_1|Y_1}(1|y_1)} \\
&= Z(X_1|Y_1).
\end{aligned}$$

Identity (a) is because Y_1 is independent of U_2 and also because Y_2 is independent of Y_1 given U_2. Identity (b) is true because X_1 is independent of Y_2U_2 given Y_1. Similarly we can easily prove that $Z^2(U_1|Y_1Y_2) = Z(X_1|Y_1)$. Now Lemma 2 implies that $Z(U_1|Y_1Y_2) \geq Z(X_1|Y_1)$. Exchanging the roles of (X_1, Y_1) and (X_2, Y_2), we also get $Z(U_1|Y_1Y_2) \geq Z(X_2|Y_2)$. Therefore $Z(U_1|Y_1Y_2) \geq \max\{Z(X_2|Y_2), Z(X_1|Y_1)\}$.We now derive the Bhattacharya parameter
$Z(U_2|Y_1Y_2U_1)$

$$\begin{aligned}
&= 2\sum_{y_1y_2u_1\in\mathcal{Y}_1\times\mathcal{Y}_2\times\mathcal{X}} P_{U_1Y_1Y_2}(u_1y_1y_2)\sqrt{P_{U_2|U_1Y_1Y_2}(0|u_1y_1y_2)P_{U_2|U_1Y_1Y_2}(1|u_1y_1y_2)} \\
&= 2\sum_{y_1y_2u_1\in\mathcal{Y}_1\times\mathcal{Y}_2\times\mathcal{X}} \frac{P_{U_1Y_1Y_2}(u_1y_1y_2)}{P_{U_1|Y_1Y_2}(u_1|y_1y_2)}\left[P_{X_1|Y_1}(u_1|y_1)P_{X_2|Y_2}(0|y_2)P_{X_1|Y_1}(u_1+1|y_1)P_{X_2|Y_2}(1|y_2)\right]^{0.5} \\
&= 2\sum_{y_1y_2u_1\in\mathcal{Y}_1\times\mathcal{Y}_2\times\mathcal{X}} P_{Y_1Y_2}(y_1y_2)\left[P_{X_1|Y_1}(u_1|y_1)P_{X_1|Y_1}(u_1+1|y_1)P_{X_2|Y_2}(0|y_2)P_{X_2|Y_2}(1|y_2)\right]^{0.5} \\
&= 2\sum_{u_1\in\mathcal{X}}\sum_{y_1\in\mathcal{Y}_1}\sum_{y_2\in\mathcal{Y}_2} P_{Y_1}(y_1)P_{Y_2}(y_2)\left[P_{X_1|Y_1}(u_1|y_1)P_{X_2|Y_2}(0|y_2)P_{X_1|Y_1}(u_1+1|y_1)P_{X_2|Y_2}(1|y_2)\right]^{0.5} \\
&= Z(X_1|Y_1)Z(X_2|Y_2).
\end{aligned}$$

□

Proof of Lemma 3:

$$||P-Q|| = \sum_{(x_1,x_2)} \frac{1}{2}|P(x_1,x_2) - Q(x_1,x_2)|$$

$$= \sum_{(x_1,x_2)} \frac{1}{2}|P_1(x_1)P_2(x_2) - Q_1(x_1)Q_2(x_2)|$$

$$= \sum_{(x_1,x_2)} \frac{1}{2}|P_1(x_1)P_2(x_2) - Q_1(x_1)P_2(x_2) + Q_1(x_1)P_2(x_2) - Q_1(x_1)Q_2(x_2)|$$

$$\overset{(a)}{\leq} \sum_{(x_1,x_2)} (\frac{1}{2}|P_1(x_1)P_2(x_2) - Q_1(x_1)P_2(x_2)| + \frac{1}{2}|Q_1(x_1)P_2(x_2) - Q_1(x_1)Q_2(x_2)|)$$

$$= \sum_{(x_1,x_2)} P_2(x_2)|P_1(x_1) - Q_1(x_1)| + \sum_{(x_1,x_2)} Q_1(x_1)|P_2(x_2) - Q_1(x_2)|$$

$$= \sum_{(x_1,x_2)} P_2(x_2)|P_1(x_1) - Q_1(x_1)| + \sum_{(x_1,x_2)} Q_1(x_1)|P_2(x_2) - Q_1(x_2)|$$

$$= \sum_{(x_1)} |P_1(x_1) - Q_1(x_1)| + \sum_{(x_2)} |P_2(x_2) - Q_1(x_2)|$$

$$= ||P_1 - Q_1|| + ||P_2 - Q_2||$$

$$= \epsilon_1 + \epsilon_2.$$

Identity (a) is true by triangular inequality. □

Proof of Lemma 4:

$$\begin{aligned} ||Q_{X,Y} - P_{X,Y}|| &= \sum_{(x,y):P(x,y)>Q(x,y)} P(x,y) - Q(x,y) \\ &= \sum_{(x,y):P(x)p(y|x)>Q(x)p(y|x)} P(x)p(y|x) - Q(x)p(y|x) \\ &= \sum_{(x,y):P(x)>Q(x)} (P(x) - Q(x))p(y|x) \\ &= \sum_{x:P(x)>Q(x)} \sum_{y} (P(x) - Q(x))p(y|x) \\ &= \sum_{x:P(x)>Q(x)} (P(x) - Q(x)) \\ &= ||Q_X - P_X||. \quad \square \end{aligned}$$

Proof of Lemma 5:

1.

The Bhattacharya parameter of the new bit-channel produced with U'_{x_j} and $U^{1:x_j-1}V^{1:y_j-1}Y^{1:2N}$

as input and output, respectively will be as follows.

$$\begin{aligned} Z(U'_{x_j}|U^{1:x_j-1}V^{1:y_j-1}Y^{1:2N}) &\overset{(a)}{\geq} \max\{Z(U_{x_j}|U^{1:x_j-1}Y^{1:N}), Z(V_{y_j}|U^{1:y_j-1}Y^{N+1:2N})\} \\ &\overset{(b)}{\geq} 1-2^{-N^{\beta}}. \end{aligned}$$

Identity (a) is true by Proposition 1. Identity (b) follows because either of $Z(U_{x_j}|U^{1:x_j-1}Y^{1:N})$ (if DMC 2 is selected in S) will be greater than $1-2^{-N^{\beta}}$ or $Z(V_{y_j}|U^{1:y_j-1}Y^{N+1:2N})$ (if DMC 1 is selected in S) will be greater than $1-2^{-N^{\beta}}$.

2.

The Bhattacharya parameter of the new bit-channel produced with V'_{y_j} and $U^{1:x_j-1}V^{1:y_j-1}U'_{x_j}Y^{1:2N}$ as input and output respectively will be as follows.

$$\begin{aligned} Z(V'_{y_j}|U^{1:x_j-1}V^{1:y_j-1}U'_{x_j}Y^{1:2N}) &\overset{(a)}{=} Z(U_{x_j}|U^{1:x_j-1}Y^{1:N})Z(V_{y_j}|V^{1:y_j-1}Y^{N+1:2N}) \\ &\leq 2^{-N^{\beta}}. \end{aligned}$$

Identity (a) is true by Proposition 1. Identity (b) follows because either of $Z(U_{x_j}|U^{1:x_j-1}Y^{1:N})$ (if DMC 1 is selected in S) will be less than $2^{-N^{\beta}}$ or $Z(V_{y_j}|U^{1:y_j-1}Y^{N+1:2N})$ (if DMC 2 is selected in S) will be less than $2^{-N^{\beta}}$. □

Proof of Theorem 2:

Let the linear bijective transform which maps $(U'^{1:N}\ \ V'^{1:N})$ to $(U^{1:N}\ \ V^{1:N})$ be H_{2N}. Let the word $(u^{1:N}\ \ v^{1:N})$ be obtained by applying H_{2N} to the word $(u'^{1:N}\ \ v'^{1:N})$. The probability that the word $(U'^{1:N}\ \ V'^{1:N})$= $(u'^{1:N}\ \ v'^{1:N})$ and received vector $Y^{1:2N}=y^{1:2N}$ when a DMC in S is selected will be as follows:

$$\begin{aligned} &2^{-(2|I_1\cap I_2|+|\mathcal{B}|)}\cdot\mathbb{1}[\cap_{i\in(\mathcal{H}_X-(I_1\cap I_2))}\{f_1(i)=u'_i\}]\cdot\mathbb{1}[\cap_{i\in(\mathcal{H}_X-((I_1\cap I_2)\cup\mathcal{B}))}\{f_2(i)=v'_i\}]\cdot \\ &\mathbb{1}[\cap_{i\in R}\{\lambda^1(u^{1:i-1})=u_i\}]\cdot\mathbb{1}[\cap_{i\in R}\{\lambda^2(v^{1:i-1})=v_i\}]\cdot\Pi_{i\in\mathcal{L}_X}(\delta_i(u_i|u^{1:i-1})\delta_i(v_i|v^{1:i-1}))\cdot \qquad (11) \\ &P_{Y^{1:N}|U^{1:N}}(y^{1:N}|u^{1:N})\cdot P_{Y^{1:N}|U^{1:N}}(y^{N+1:2N}|v^{1:N}). \end{aligned}$$

Note that we used the fact, $P_{U^{1:N}Y^{1:N}}=P_{V^{1:N}Y^{N+1:2N}}$. Here the measure P is induced by equation (1). Let $\mathcal{E}_i^b$ be the error event for the ith bit channel of block b.

For $i\in I_1\cap I_2$, we define the following events:

$$\mathcal{E}_i^1=\{(u'^{1:N},v'^{1:N},y^{1:2N}):P_{U_i|U^{1:i-1}Y^{1:N}}(u_i+1|u^{1:i-1}y^{1:N})\geq P_{U_i|U^{1:i-1}Y^{1:N}}(u_i|u^{1:i-1}y^{1:N})\},$$

$$\mathcal{E}_i^2=\{(u'^{1:N},v'^{1:N},y^{1:2N}):P_{U_i|U^{1:i-1}Y^{1:N}}(v_i+1|v^{1:i-1}y^{1:N})\geq P_{U_i|U^{1:i-1}Y^{1:N}}(v_i|v^{1:i-1}y^{1:N})\}.$$

For $j \in [G]$, we define the following event:

$$\begin{aligned}\mathcal{E}^2_{y_j} = \{(u'^{1:N}, v'^{1:N}, y^{1:2N}) : &P_{V'_{y_j}|U^{1:x_j-1}U'_{x_j}V^{1:y_j-1}Y^{1:2N}}(u_i + 1|u^{1:x_j-1}u'_{x_j}v^{1:y_j-1}y^{1:2N}) \\ &\geq P_{V'_{y_j}|U^{1:x_j-1}U'_{x_j}V^{1:y_j-1}Y^{1:2N}}(u_i|u^{1:x_j-1}u'_{x_j}v^{1:y_j-1}y^{1:2N})\}.\end{aligned} \tag{12}$$

Therefore the error event $\mathcal{E}$ becomes as follows:

$$\mathcal{E} = \{\cup_{i \in I_1 \cap I_2} \mathcal{E}^1_i\} \cup \{\cup_{i \in (I_1 \cap I_2) \cup \mathcal{B}} \mathcal{E}^2_i\}.$$

Now the probability of error for the given $f_1, f_2, \lambda^1_R, \lambda^2_R$ will be as follows:

$$\begin{aligned}&P_{e,l}(\lambda^1_R, \lambda^2_R, f_1, f_2) \\ &\quad= \sum_{(u'^{1:N}, v'^{1:N}, y^{1:2N})} 2^{-(2|I_1 \cup I_2| + |\mathcal{B}|)} \cdot \mathbb{1}[\cap_{i \in (\mathcal{H}_X - (I_1 \cap I_2))}\{f_1(i) = u'_i\}] \cdot \\ &\qquad\qquad \mathbb{1}[\cap_{i \in (\mathcal{H}_X - ((I_1 \cap I_2) \cup \mathcal{B}))}\{f_2(i) = v'_i\}] \cdot \mathbb{1}[\cap_{i \in R}\{\lambda^1(u^{1:i-1}) = u_i\}] \cdot \\ &\qquad\qquad \mathbb{1}[\cap_{i \in R}\{\lambda^2(v^{1:i-1}) = v_i\}] \cdot \Pi_{i \in \mathcal{L}_X}(\delta_i(u_i|u^{1:i-1})\delta_i(v_i|v^{1:i-1})) \cdot \\ &\qquad\qquad P_{Y^{1:N}|U^{1:N}}(y^{1:N}|u^{1:N})P_{Y^{1:N}|U^{1:N}}(y^{N+1:2N}|v^{1:N}) \cdot \\ &\qquad\qquad \mathbb{1}[(u'^{1:N}, v'^{1:N}, y^{1:2N}) \in \mathcal{E}].\end{aligned} \tag{13}$$

Now we take the ensemble expectation of $P_{e,l}$,

$$\begin{aligned}\mathbb{E}[P_{e,l}(\lambda^1_R, \lambda^2_R, f_1, f_2)] = \sum_{(u'^{1:N}, v'^{1:N}, y^{1:2N})} &2^{-2|\mathcal{H}_X|}\Pi_{i \in \mathcal{L}_X}(\delta_i(u_i|u^{1:i-1}) \\ &\delta_i(v_i|v^{1:i-1}))\Pi_{i \in R}(P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}) \\ &P_{U_i|U^{1:i-1}}(v_i|v^{1:i-1}))P_{Y^{1:N}|U^{1:N}}(y^{1:N}|u^{1:N}) \\ &P_{Y^{1:N}|U^{1:N}}(y^{N+1:2N}|v^{1:N})\mathbb{1}[(u'^{1:N}, v'^{1:N}, y^{1:2N}) \in \mathcal{E}].\end{aligned} \tag{14}$$

Now we define the measure Q by considering the random variables $U'^{1:N}V'^{1:N}Y^{1:2N}$ as follows:

$$\begin{aligned}Q_{U'^{1:N}V'^{1:N}Y^{1:2N}}(u'^{1:N}, v'^{1:N}, y^{1:2N}) :\overset{(a)}{=} &2^{-2|\mathcal{H}_X|}\Pi_{i \in \mathcal{L}_X}(\delta_i(u_i|u^{1:i-1})\delta_i(v_i|v^{1:i-1})) \\ &\Pi_{i \in R}(P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}) \\ &P_{U_i|U^{1:i-1}}(v_i|v^{1:i-1}))P_{Y^{1:N}|U^{1:N}}(y^{1:N}|u^{1:N}) \\ &P_{Y^{1:N}|U^{1:N}}(y^{N+1:2N}|v^{1:N}).\end{aligned} \tag{15}$$

Identity (a) is true since $(u^{1:N} \quad v^{1:N})$ is obtained by applying H_{2N} map on the word $(u'^{1:N} \quad v'^{1:N})$. Note that

$$Q_{U'^{1:N}V'^{1:N}Y^{1:2N}}(u'^{1:N}, v'^{1:N}, y^{1:2N}) = Q_{U^{1:N}V^{1:N}Y^{1:2N}}(u^{1:N}, v^{1:N}, y^{1:2N}).$$

From equations (14) and (15), we have the following:

$$Q_{U'^{1:N}V'^{1:N}Y^{1:2N}}(\mathcal{E}) = \mathbb{E}[P_{e,l}(\lambda_R^1, \lambda_R^2, f_1, f_2)]. \tag{16}$$

By marginalizing the distribution in the equation (15) over the random variables $(V^{1:N}, Y^{N+1:2N})$ and $(U^{1:N}, Y^{1:N})$, respectively we will have the following:

$$\begin{gathered} Q_{U^{1:N}Y^{1:N}}(u^{1:N}, y^{1:N}) = 2^{-|\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X}\delta_i(u_i|u^{1:i-1})\Pi_{i\in R}P_{U_i|U^{1:i-1}}(u_i|u^{1:i-1}) \\ P_{Y^{1:N}|U^{1:N}}(y^{1:N}|u^{1:N}). \end{gathered} \tag{17}$$

$$\begin{gathered} Q_{V^{1:N}Y^{N+1:2N}}(v^{1:N}, y^{N+1:2N}) = 2^{-|\mathcal{H}_X|}\Pi_{i\in\mathcal{L}_X}\delta_i(v_i|v^{1:i-1})\Pi_{i\in R}P_{U_i|U^{1:i-1}}(v_i|v^{1:i-1}) \\ P_{Y^{1:N}|U^{1:N}}(y^{N+1:2N}|v^{1:N}). \end{gathered} \tag{18}$$

Note that,

$$Q_{U^{1:N}V^{1:N}Y^{1:2N}}(u^{1:N}, v^{1:N}, y^{1:N}) = Q_{U^{1:N}Y^{1:N}}(u^{1:N}, y^{1:N}).Q_{V^{1:N}Y^{N+1:2N}}(v^{1:N}, y^{N+1:2N}). \tag{19}$$

Therefore $(U^{1:N}, Y^{1:N})$ and $(V^{1:N}, Y^{N+1:2N})$ are independent and identically distributed with respect to measure Q. So using $Q_{U^{1:N}Y^{1:N}} = Q_{V^{1:N}Y^{N+1:2N}}$ and also equation (1) we get

$$||Q_{U^{1:N}Y^{1:N}} - P_{U^{1:N}Y^{1:N}}|| = ||Q_{V^{1:N}Y^{N+1:2N}} - P_{V^{1:N}Y^{N+1:2N}}||. \tag{20}$$

Now we bound the probability of error as follows:

$$\begin{aligned} Q_{U'^{1:N}V'^{1:N}Y^{1:2N}}(\mathcal{E}) &\le ||Q_{U'^{1:N}V'^{1:N}Y^{1:2N}} - P_{U'^{1:N}V'^{1:N}Y^{1:2N}}|| + P_{U'^{1:N}V'^{1:N}Y^{1:2N}}(\mathcal{E}) \\ &\le ||Q_{U^{1:N}V^{1:N}Y^{1:2N}} - P_{U^{1:N}V^{1:N}Y^{1:2N}}|| + P_{U^{1:N}V^{1:N}Y^{1:2N}}(\mathcal{E}) \\ &\le ||Q_{U^{1:N}V^{1:N}Y^{1:2N}} - P_{U^{1:N}V^{1:N}Y^{1:2N}}|| + \sum_{i\in I_1\cap I_2} P_{U'^{1:N}V'^{1:N}Y^{1:2N}}(\mathcal{E}_i^1) \\ &\quad + \sum_{i\in(I_1\cap I_2)\cup\mathcal{B}} P_{U'^{1:N}V'^{1:N}Y^{1:2N}}(\mathcal{E}_i^2). \end{aligned} \tag{21}$$

Now we bound each term in the last inequality of the summation. We now consider the first term of the summation.

$$\begin{aligned} &||Q_{U^{1:N}V^{1:N}Y^{1:2N}} - P_{U^{1:N}V^{1:N}Y^{1:2N}}|| \\ &\quad= \frac{1}{2}\sum_{(u^{1:N},v^{1:N},y^{1:2N})} |Q_{U^{1:N}V^{1:N}Y^{1:2N}}(u^{1:N}, v^{1:N}, y^{1:2N}) - P_{U^{1:N}V^{1:N}Y^{1:2N}}(u^{1:N}, v^{1:N}, y^{1:2N})| \\ &\quad= \frac{1}{2}\sum_{(u^{1:N},v^{1:N},y^{1:2N})} |Q_{U^{1:N}Y^{1:N}}(u^{1:N}, y^{1:N})Q_{V^{1:N}Y^{N+1:2N}}(v^{1:N}, y^{N+1:2N}) - \\ &\qquad\qquad P_{U^{1:N}Y^{1:N}}(u^{1:N}, y^{1:N})P_{V^{1:N}Y^{N+1:2N}}(v^{1:N}, y^{N+1:2N})| \end{aligned}$$

Now we upper bound it as follows:

$$
\begin{aligned}
&||Q_{U^{1:N}V^{1:N}Y^{1:2N}} - P_{U^{1:N}V^{1:N}Y^{1:2N}}||\\
&\quad = \frac{1}{2}\sum_{(u^{1:N},v^{1:N},y^{1:2N})} |Q_{U^{1:N}Y^{1:N}}(u^{1:N},y^{1:N})Q_{V^{1:N}Y^{N+1:2N}}(v^{1:N},y^{N+1:2N}) -\\
&\qquad Q_{U^{1:N}Y^{1:N}}(u^{1:N},y^{1:N})P_{V^{1:N}Y^{N+1:2N}}(v^{1:N},y^{N+1:2N}) +\\
&\qquad Q_{U^{1:N}Y^{1:N}}(u^{1:N},y^{1:N})P_{V^{1:N}Y^{N+1:2N}}(v^{1:N},y^{N+1:2N}) -\\
&\qquad P_{U^{1:N}Y^{1:N}}(u^{1:N},y^{1:N})P_{V^{1:N}Y^{N+1:2N}}(v^{1:N},y^{N+1:2N})|\\
&\quad \overset{(a)}{\leq} \frac{1}{2}\sum_{(u^{1:N},v^{1:N},y^{1:2N})} Q_{U^{1:N}Y^{1:N}}(u^{1:N},y^{1:N})|(Q_{V^{1:N}Y^{N+1:2N}}(v^{1:N},y^{N+1:2N}) -\\
&\qquad P_{V^{1:N}Y^{N+1:2N}}(v^{1:N},y^{N+1:2N}))| + |(Q_{U^{1:N}Y^{1:N}}(u^{1:N},y^{1:N}) -\\
&\qquad P_{U^{1:N}Y^{1:N}}(u^{1:N},y^{1:N}))|P_{V^{1:N}Y^{N+1:2N}}(v^{1:N},y^{N+1:2N})\\
&\quad = \frac{1}{2}\sum_{(v^{1:N},y^{N+1:2N})} |Q_{V^{1:N}Y^{N+1:2N}}(v^{1:N},y^{N+1:2N}) - P_{V^{1:N}Y^{N+1:2N}}(v^{1:N},y^{N+1:2N})|\\
&\qquad + \frac{1}{2}\sum_{(u^{1:N},y^{1:N})} |Q_{U^{1:N}Y^{1:N}}(u^{1:N},y^{1:N}) - P_{U^{1:N}Y^{1:N}}(u^{1:N},y^{1:N})| \qquad (22)\\
&\quad = ||Q_{V^{1:N}Y^{N+1:2N}} - P_{V^{1:N}Y^{N+1:2N}}|| + ||Q_{U^{1:N}Y^{1:N}} - P_{U^{1:N}Y^{1:N}}||\\
&\quad \overset{(b)}{=} 2||Q_{U^{1:N}Y^{1:N}} - P_{U^{1:N}Y^{1:N}}|| = O(2^{-N^{\beta}}).
\end{aligned}
$$

Identity (a) is true by triangular inequality. Identity (b) is true from equation (20). The above steps are equivalent to applying Lemma 3.

For $i \in I_1 \cap I_2$, $P(\mathcal{E}_i^b)$ can be bounded as in equation (60) in [7] which will become as follows:

$$
P(\mathcal{E}_i^b) \leq 2^{-N^{\beta}}. \qquad (23)
$$

For $i \in \mathcal{B}$ there exists a $j \in [G]$ such that $i = y_j$. We upperbound $P(\mathcal{E}_i^2)$ corresponding to block 2 as follows:

$$
\begin{aligned}
P_{U'^{1:N},V'^{1:N}Y^{1:2N}}(\mathcal{E}_i^2) \leq &\sum_{u'^{1:x_j}v'^{1:y_i-1}y^{1:N}} P(u'^{1:x_j}v'^{1:y_j-1},y^{1:2N})P(v'_{y_j}|u'^{1:x_j}v'^{1:y_j-1},y^{1:2N})\\
&1[P(v'_{y_j}|u'^{1:x_j}v'^{1:y_j-1},y^{1:2N}) \leq P(v'_{y_j}+1|u'^{1:x_j}v'^{1:y_j-1},y^{1:2N})]\\
\leq &\sum_{u'^{1:x_j}v'^{1:y_i-1}y^{1:N}} P(u'^{1:x_j}v'^{1:y_j-1},y^{1:2N})P(v'_{y_j}|u'^{1:x_j}v'^{1:y_j-1},y^{1:2N})
\end{aligned}
$$

$$\sqrt{\frac{P(v'_{y_j}+1|u'^{1:x_j}v'^{1:y_j-1},y^{1:2N})}{P(v'_{y_j}|u'^{1:x_j}v'^{1:y_j-1},y^{1:2N})}}$$

Therefore,

$$\begin{aligned} P_{U'^{1:N},V'^{1:N}Y^{1:2N}}(\mathcal{E}_i^2) &\leq Z(V'_{y_j}|U'^{1:x_j-1}V'^{1:y_j-1}Y^{1:2N}) \\ &\overset{(a)}{\leq} 2^{-N^{\beta'}}. \end{aligned} \tag{24}$$

Identity (a) is true by Lemma 5.

From equations (21), (22), (23) and (24), we conclude that

$$\mathbb{E}[P_{e,l}(\lambda_R^1,\lambda_R^2,f_1,f_2)] = O(2^{-N^{\beta'}})$$

for each $l \in S = \{1, 2\}$. □