2019 57th Annual Allerton Conference on Communication, Control, and Computing
(Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, September 24-27, 2019

Slepian-Wolf Polar Coding with Unknown Correlation

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

***Abstract*—We consider the source coding problem of a binary discrete memoryless source with correlated side information available only at the receiver whose conditional distribution given the source is unknown to the encoder. We propose two methods based on polar codes to attain the achievable rates under this setting. The first method incorporates a staircase scheme, which has been used for universal polar coding for a compound channel. The second method is based on the technique of universalization using bit-channel combining. We also give a list of pros and cons for the two proposed methods.**

I. Introduction

A. Background

Arikan [1] constructed capacity-achieving codes for binary-input symmetric channels. Korada and Urbanke [7] constructed a Slepian-Wolf polar coding scheme for two correlated sources under some assumptions. Arikan [2] proposed a polar coding method for an arbitrary discrete memoryless source with correlated side-information available at the receiver. From there, he also derived a Slepian-Wolf polar coding strategy for any two binary correlated random variables. Arikan [3] proposed a monotone chain rule to achieve all the rates of the Slepian-Wolf region without the use of a time-sharing policy.

A capacity-achieving coding scheme based on source and channel polarization for binary-input asymmetric channels was proposed by Honda and Yamamoto [6]. Hassani and Urbanke [4], [5] presented universal coding schemes to achieve rates close to the compound capacity of binary-input symmetric discrete-memoryless channels (DMCs) that are based on polar codes. The authors [8] proposed a universal polar coding scheme for the asymmetric setting that eliminates the need to store high complexity boolean functions. The scheme uses the elements of coding strategies in [4], [6]. Wang and Kim [11] discussed the linear code duality between channel coding and source coding when the correlated side information is available at the receiver. In this paper, we consider the variant of the Slepian-Wolf coding problem which involves a binary memoryless source and correlated side information available at the receiver, as usual, but where the conditional distribution of the side information given the source is unknown to the encoder.

B. Problem definition

Let $\mathcal{X}$ be the binary alphabet and $\mathcal{Y}$ be some arbitrary finite alphabet. A binary discrete memoryless source X_i is distributed as $P_X(x)$ with side information Y_i available at the receiver. The $(X_i, Y_i)_{i=1}^{\infty}$ sequence is an iid random process whose joint distribution is $P_X(x)p(y|x)$. The conditional distribution $p(y|x)$ is unknown to the encoder, but available to the decoder only. We also assume that $p(y|x)$ is known to come from a class $\mathcal{C}$ of conditional distributions of a random variable over the alphabet $\mathcal{Y}$ given a correlated random variable over $\mathcal{X}$. The class $\mathcal{C}$ is available to the encoder.
A $(2^k, n)$ code for the defined problem consists of

- an encoder $g : \mathcal{X}^n \to \{1 : 2^k\}$, and
- a decoder $h : \{1 : 2^k\} \times \mathcal{Y}^n \to \mathcal{X}^n$

where n is the block length and $\frac{k}{n}$ is called the rate of the code. Let $P_e^{(n)} = P(X^n \neq h(g(X^n), Y^n))$ be the probability of error. If there is a sequence of $(2^{nR}, n)$ codes for which the corresponding sequence of $P_e^{(n)}$ goes to zero, then the rate R is achieved. Note that classical Slepian-Wolf coding is the case when $p(y|x)$ is known to both the encoder and decoder. In that case, we know that the rate R is achieved if and only if $R > H(X|Y)$. Therefore the achievable rates of the proposed problem should be greater than $\max_{p(y|x)\in\mathcal{C}} H(X|Y)$ where (X, Y) is distributed as $P_X(x)p(y|x)$.

C. Contribution

We derive two coding strategies for the proposed setting based on the universal polar coding schemes for a compound channel [5], [8]. This will establish the duality between the coding strategies for these source and channel coding settings. Our first method can achieve all rates greater than $\max_{p(y|x)\in\mathcal{C}} H(X|Y)$ for a uniformly distributed source when the class $\mathcal{C}$ contains only conditional distributions with properties of a symmetric channel. The second method can achieve all rates greater than $\max_{p(y|x)\in\mathcal{C}} H(X|Y)$ when $\mathcal{C}$ is a finite set for any arbitrary source.

In Section II, we introduce some definitions and notation which will be used throughout the paper. In Section III, we describe the Slepian-Wolf polar coding with correlated side information available at the receiver. In Section IV, we describe our first method that uses the idea of the staircase scheme for a uniform source. We also give the coding strategy for a non-uniformly distributed source. In Section V, we explain the second method which is based on the technique of universalization using bit-channel combining.

978-1-7281-3151-1/19/$31.00 ©2019 IEEE

II. Preliminaries

Definition 1. *A binary-input discrete memoryless channel with output alphabet $\mathcal{Y}$ with transition probabilities $p(y|x)$ for each $(x, y) \in \{0, 1\} \times \mathcal{Y}$ is said to be symmetric if there exists a permutation $\pi_1 : \mathcal{Y} \to \mathcal{Y}$ such that $\pi_1 = \pi_1^{-1}$ and $p(y|x) = p(\pi_a(y)|x + a)$ for each $(x, a, y) \in \{0, 1\}^2 \times \mathcal{Y}$, where $\pi_o : \mathcal{Y} \to \mathcal{Y}$ is the identity permutation.*

We denote the row vector $(\pi_{s_1}(y_1), \pi_{s_2}(y_2), ..., \pi_{s_N}(y_N))$ as $s^{1:N}.y^{1:N}$ for any $y^{1:N} \in \mathcal{Y}^N$ and $s^{1:N} \in \{0, 1\}^N$, where $\pi_0 : \mathcal{Y} \to \mathcal{Y}$ is the identity permutation and $\pi_1 : \mathcal{Y} \to \mathcal{Y}$ is the permutation corresponding to a symmetric channel.

Let G_N be the conventional polar transform, represented by a binary matrix of dimension $N \times N$. If $U^{1:N} = X^{1:N} G_N$, then we denote $P(U^{1:N} = u^{1:N})$ by $P_{U^{1:N}}(u^{1:N})$ and similarly we denote $P(U_i = u_i | U^{1:i-1} Y^{1:N} = u^{1:i-1} y^{1:N})$ by $P_{U_i|U^{1:i-1}Y^{1:N}}(u_i|u^{1:i-1}y^{1:N})$. We denote the subvector of $U^{1:N}$ corresponding to the bit-channel set $\mathcal{A} \subset \{1 : N\}$ as $U^{\mathcal{A}}$.

Let $\mathcal{C} = \{p_1(y|x), p_2(y|x), ..., p_s(y|x)\}$, $s \in \mathbb{N}$. Let $(X_i, Y_i)_{i=1}^N$ be iid random tuples distributed according to $P_X(x) p_l(y|x)$, where $l \in \{1 : s\}$ and $N = 2^n$. For the random variable pair (X, Y) distributed as $P_X(x) p_l(y|x)$, the Bhattacharyya parameter is defined as

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

We define the following bit-channel subsets as follows, where $\beta < 0.5$.

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

Note that $\mathcal{L}_X \subseteq \mathcal{L}_{X|Y_l}$, for each $l \in \{1 : s\}$. We have the following results from Theorem 1 in [6].

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

We remove the subscript l for denoting the bit-channel sets $\mathcal{L}_{X|Y_l}$ and $\mathcal{H}_{X|Y_l}$ whenever (X, Y) is distributed as $P_X(x) p(y|x)$ and denote them as $\mathcal{L}_{X|Y}$ and $\mathcal{H}_{X|Y}$, respectively.

Let the $p(y|x)$s for each $(x, y) \in \mathcal{X} \times \mathcal{Y}$ be the transition probabilities of a symmetric channel. Let $(X_i, Y_i)_{i=1}^N$ be iid random variable pairs distributed according to $P_X(x) p(y|x)$ where $P_X(x)$ is distributed as Bern($\frac{1}{2}$). Let $U^{1:N} = X^{1:N} G_N$. Then the MAP (ML) decision rule for the bit-channel $i \in \{1 : N\}$ in this setting will be the function $\Phi_i : \{0, 1\}^{i-1} \times \mathcal{Y}^N \to \{0, 1\}$ defined as follows.

$$\Phi_i(u^{1:i-1}, y^{1:N}) = \mathbb{1}\{P_{U^{1:i-1}, Y^{1:N}|U_i}(\hat{u}^{1:i-1}, y^{1:N}|1) \geq P_{U^{1:i-1}, Y^{1:N}|U_i}(\hat{u}^{1:i-1}, y^{1:N}|0)\}.$$

Φ_i is precisely the decision rule used in the successive cancellation (SC) decoding for the bit-channel $i \in \mathcal{L}_{X|Y}$ in the polar code construction for symmetric channels. Let us denote the Bhattacharyya parameter corresponding to the bit-channel $i \in \{1 : N\}$ as Z_i. Therefore $Z_i = Z(U_i|U^{1:i-1} Y^{1:N})$ in this setting.

III. Source coding with side-information (Slepian-Wolf Polar coding)

We revisit the polar coding scheme proposed by Arikan [2] for the Slepian-Wolf setting that has the binary discrete memoryless source X_i distributed as $P_X(x)$ with correlated side information Y_i available at the receiver, $i \in \{1 : N\}$. $(X_i, Y_i)_{i=1}^N$ is an iid process whose joint distribution is $P_X(x) p(y|x)$. Here we assume that $p(y|x)$ is known to both the encoder and decoder. The encoding algorithm is presented below.

```
Encoding
Input: $X^{1:N}$ source sequence.
Output: Compressed bit stream corresponding to the source sequence.
  • Compute $U^{1:N} = X^{1:N} G_N$.
  • Transmit $U^{\mathcal{L}^c_{X|Y}}$.
```

The decoding method is as follows.

```
Decoding
Input: Correlated side information $Y^{1:N}$ and $U^{\mathcal{L}^c_{X|Y}}$.
Output: Source estimate $\hat{X}^{1:N}$.
for $i = 1 : N$
    1. If $i \in \mathcal{L}^c_{X|Y}$, set $\hat{U}_i = U_i$.
    2. If $i \in \mathcal{L}_{X|Y}$, set
       $\hat{U}_i = \mathbb{1}\{P_{U_i|U^{1:i-1}, Y^{1:N}}(1|\hat{U}^{1:i-1}, Y^{1:N}) \geq P_{U_i|U^{1:i-1}, Y^{1:N}}(0|\hat{U}^{1:i-1}, Y^{1:N})\}$.
end
Decode $\hat{X}^{1:N}$ as $\hat{U}^{1:N} G_N$.
```

Note that the conditional distribution $P_{U_i|U^{1:i-1}, Y^{1:N}}(.|.)$ used above in the decoding algorithm is derived under the setting where $X^{1:N} = U^{1:N} G_N$ and $(X_i, Y_i)_{i=1}^N$ is iid distributed as $P_X(x) p(y|x)$. Arikan [2] proved that the probability of error for this scheme is $O(2^{N^{-\beta}})$ where $\beta < 0.5$. In our setup, however, the actual conditional distribution $p(y|x)$ is unknown to the encoder. The encoder only knows that the conditional distribution is selected from the class $\mathcal{C}$. If the encoder transmits $U^{(\cap_{i \in \mathcal{C}} \mathcal{L}_{X|Y})^c}$, then the

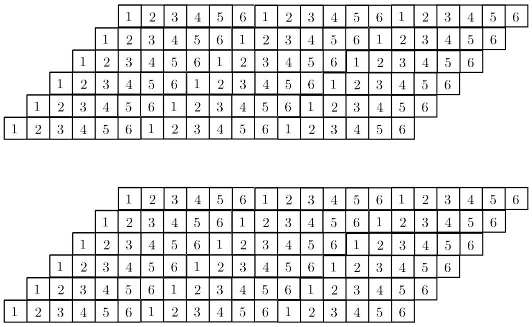

Fig. 1. Staircase with k=3, N=6 and q=2

decoder can reliably decode the bits corresponding to bit-channels $(\cap_{i\in\mathcal{C}}\mathcal{L}_{X|Y})$. However, the fraction of bit-channels $(\cap_{i\in\mathcal{C}}\mathcal{L}_{X|Y})^c$ with respect to the block length may not be going to $\max_{p(y|x)\in\mathcal{C}} H(X|Y)$ as block length grows and the fraction may always be larger than $\max_{p(y|x)\in\mathcal{C}} H(X|Y)$ by at least some positive constant. The following sections provide the source coding methods that can guarantee any rate greater than $\max_{p(y|x)\in\mathcal{C}} H(X|Y)$.

IV. Staircase scheme

In this section, we assume that the source is a binary symmetric source (uniform) and all the condition distributions in $\mathcal{C}$ are of symmetric channel type.

A. Code construction

To construct the staircase, we consider polar blocks of size N, where N is sufficiently large for polarization so that we get the $H(X|Y)$ fraction of bad bit-channels ($\mathcal{H}_{X|Y}$) and $1 - H(X|Y)$ fraction of good bit-channels ($\mathcal{L}_{X|Y}$) for each $p(y|x) \in \mathcal{C}$. We need an MDS code in the code construction. We use a Reed-Solomon code of block length $2^q - 1$ over a field GF(2^q) as an MDS code where q is $\log_2(N)$. We consider the RS code that has $L = \min_{p(y|x)\in\mathcal{C}} |\mathcal{L}_{X|Y}| - 1$ information bits. Let $\mathcal{M}$ be the set of codewords of such an RS code.

We arrange N polar blocks of size N one above the other like a staircase which will be of height N. We extend the staircase by placing $k \in \mathbb{N}$ such staircases side-by-side. Now place q such extended staircases, one above the other. So the total number of polar blocks would be Nqk. This is illustrated in Fig. 1 for the case $N = 6$, $k = 3$, and $q = 2$.

A staircase scheme designed for a compound channel with the class $\mathcal{C}$ of symmetric channels [4] is a binary linear code after all. A naive Slepian-Wolf code derived using the method [11] requires the computation of high dimensional systematic parity-check matrix ($q[(N - L)(1 + N(k - 1)) + N(N - 1)] \times N^2qk$) for the universal channel code. We avoid the computation of such a high dimensional parity-check matrix and its use in our staircase code construction. We can also get a delay saving by continuous, sequential encoding and decoding of substaircases in our staircase implementation, similar to universal channel coding [9].

While encoding, we do the compression for all the polar blocks column-by-column from left to right in the staircase structure, and we follow the same order for decoding. So, we deal with the bit-channels of 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. Now we describe the encoding algorithm for the compression.

Encoding
Input: $X^{1:N}$ source sequence corresponding to each polar block in all q staircases.
Output: Compressed bit stream for all the columns.

- Compute $U^{1:N} = X^{1:N}G_N$ for each polar block in all the q staircases.
- Now we start encoding the non-full-height columns on the left.
  - The U_is of the non-full-height columns on the left side are transmitted as is without any further compression for all q staircases. Note that this will not affect the rate as the fraction of these columns is diminishing as k approaches infinity.
- Next we start encoding the full-height columns $t = N \leq i \leq kN$.
  - In the full-height column t, there is a polar block corresponding to each index $i \in \{1 : N\}$ in all q staircases.
  - In the column t, for each $i \in \{1 : N\}$, the U_is of corresponding polar blocks for all q staircases will be read. Those q bits corresponding to the index $i \in \{1 : N\}$ will be the binary representation of a field element in GF(2^q). Hence we can read N finite field elements in the column. Let us call this vector $V^{1:N}$ in GF(2^q).
  - $V^{1:N-1}$ will be decomposed as an RS codeword and the error vector in a unique way using a systematic encoding method for the RS code.
  - We designate the positions $1 : L$ for information symbols. So, we generate the codeword in $\mathcal{M}$ corresponding to data $V^{1:L}$ by systematic encoding. The parity symbols in the systematic encoding method for the RS code can be computed by determining a remainder of a polynomial. This can be implemented using a shift register circuit with multipliers and adders [10].
  - Let the encoded codeword be $V'^{1:N-1}$. Now the error vector will be $E^{1:N-1} = V^{1:N-1}$ - $V'^{1:N-1}$. Note that $E^{1:L}$ will be zero always. We set the Nth position of the error vector $E^{1:N}$, $E_N = V_N$.
  - We transmit $E^{L+1:N}$ in the binary representation.
  - The error vector $E^{1:N}$ can also be generated by computing the syndrome of $V^{1:N}$ using the systematic parity-check matrix. This is shown in Lemma 1.

However we propose to use the shift register circuit implementation to get the systematic RS codeword [10] without explicitly computing the systematic generator or parity-check matrix.

  - This decomposition is also equivalent to standard array decoding with coset leaders of the form $[0^{1:L}, x^{L+1:N-1}]$ where $x^{L+1:N-1}$ is a vector in GF(2^q).

- Now we start encoding the non-full-height columns on the right.
  - The U_is of the non-full-height columns on the right side are transmitted as is without any further compression for all q staircases. Note that this will not affect the rate as these columns are diminishing in fraction as k approaches ∞.

Lemma 1. *Let $V^{1:N-1}$ be any $N-1$ dimensional vector in GF(2^q). Let $V'^{1:N-1}$ be the Reed-Solomon codeword in $\mathcal{M}$ corresponding to the data symbol stream $V^{1:L}$ in the systematic representation. Let $E^{1:N-1}$ be the error $V^{1:N-1} - V'^{1:N-1}$. The syndrome of the word $V^{1:N-1}$ when computed with the systematic parity-check matrix becomes $E^{L+1:N-1}$.*

Proof: Let the systematic parity-check matrix be

$$H_{sys} = \begin{bmatrix} A & I \end{bmatrix}$$

where A is a $N-1-L \times L$ dimensional matrix in GF(2^q) and I is the $N-1-L \times N-1-L$ dimensional identity matrix. Then,

$$\begin{aligned} H_{sys}(V^{1:N-1})^T &= H_{sys}((V'^{1:N-1})^T + (E^{1:N-1})^T) \\ &\overset{(a)}{=} H_{sys}(E^{1:N-1})^T \\ &= \begin{bmatrix} A & I \end{bmatrix} (E^{1:N-1})^T \\ &\overset{(b)}{=} (E^{L+1:N})^T. \end{aligned}$$

We get the identity (a) because $V'^{1:N-1}$ is a codeword in $\mathcal{M}$. So its multiplication with systematic parity check matrix should be zero. Identity (b) follows because $E^{1:L}$ is a zero vector. □

Before turning to the decoding algorithm, let us define

$$U'^{1:N} := U^{1:N} - E'^{1:N}$$

for each polar block in all q staircases, where $E'^{1:N}$ is the horizontal error vector computed for each polar block in all q staircases by transforming the vertical error vector $E^{1:N}$ corresponding to each full-height column. So we have,

$$U^{1:N}G_N = U'^{1:N}G_N + E'^{1:N}G_N.$$

That implies,

$$X^{1:N} = U'^{1:N}G_N + S^{1:N}.$$

where $S^{1:N} = E'^{1:N}G_N$ for each polar block in all q staircases. Note that we are transmitting $N-L$ bits for each full-height column. The rate for each full-height column is $\frac{N-L}{N}$, which can be made arbitrarily close to $\max_{p(y|x)\in\mathcal{C}} H(X|Y)$ for sufficiently large N. We did not compress the bit-stream corresponding to the non-full-height columns, but their effect on the overall rate can be made arbitrarily small for a sufficiently large k because as k goes to ∞, the fraction of the number of bits in the non-full-height columns with respect to total block length goes to zero.

Lemma 2. *Let $u'^{1:N}$ and $s^{1:N}$ be any two binary vectors. The conditional distribution of permuted side information $s^{1:N}.Y^{1:N}$ given $X^{1:N} = u'^{1:N}G_N + s^{1:N}$ will be the same as the conditional distribution of the received vector given the word $u'^{1:N}G_N$ is transmitted over the symmetric channel $p(y|x)$.*

Proof:

$$\begin{aligned} &P(s^{1:N}.Y^{1:N} = y^{1:N} | X^{1:N} = u'^{1:N}G_N + s^{1:N}) \\ &= P(Y^{1:N} = s^{1:N}.y^{1:N} | X^{1:N} = u'^{1:N}G_N + s^{1:N}) \\ &= \Pi_{i\in 1:N} p(s_i.y_i|(u'^{1:N}G_N)_i + s_i) \\ &= \Pi_{i\in 1:N} p(s_i.y_i|(u'^{1:N}G_N)_i + s_i) \\ &\overset{(a)}{=} \Pi_{i\in 1:N} p(s_i.s_i.y_i|(u'^{1:N}G_N)_i) \\ &= \Pi_{i\in 1:N} p(y_i|(u'^{1:N}G_N)_i). \end{aligned}$$

The identity (a) follows from the symmetric channel property. The term $\Pi_{i\in 1:N} p(y_i|(u'^{1:N}G_N)_i)$ is precisely the conditional probability of getting the received vector $y^{1:N}$ given the word $u'^{1:N}G_N$ is transmitted over the symmetric channel $p(y|x)$. This concludes the proof. □

Fact 1. *Let $u'^{1:N}$ be any binary vector. The conditional probability of error of all the bit-channels in $\mathcal{L}_{X|Y}$ is upper bounded by $2^{-N^{\beta}}$ given that $u'^{1:N}G_N$ is transmitted over the symmetric channel $p(y|x)$ for any $\beta < 0.5$.*

The above fact follows from Arikan's capacity-achieving polar coding construction for symmetric channels [1] where it was proved that the conditional probability of error of a bit-channel given that any particular word is transmitted over the channel is always the same irrespective of the word that is transmitted. Now we start describing the decoding algorithm.

Decoding
Input: Side information $Y^{1:N}$ for each block and $E^{L+1:N}$ for each full-height column.
Output: Estimates of $X^{1:N}$ corresponding to all polar blocks.

- Using error vectors $E^{L+1:N}$ corresponding to all full-height columns, the decoder computes the horizontal error vectors $E'^{1:N}$ corresponding to all polar blocks in all q staircases.
- Now the decoder estimates U_i's of each column that corresponds to different polar blocks from left to right. Then the estimation of $U'^{1:N}$ leads to estimate $U^{1:N}$ by adding $E'^{1:N}$ for all polar blocks in all q staircases . Let the estimates be denoted as $\hat{U}'^{1:N}$, $\hat{U}^{1:N}$ and $\hat{X}^{1:N}$.

- Decoding the non-full-height columns on the left side.
  - The U_is corresponding to these columns are transmitted as is by the encoder.
  - Hence $\hat{U}_i' = U_i - E_i' = U_i$ for all these columns in all q staircases.
- To decode full-height columns from $t = N \leq i \leq kN$:
  - The decoder has knowledge of the exact $p(y|x) \in \mathcal{C}$. For the blocks corresponding to bit-channels $i \in \mathcal{L}_{X|Y}$, we use the following decision rule to decode U_i's. The idea follows from Lemma 2 and Fact 1.

$$\hat{U}_i' = \Phi_i(\hat{U}'^{1:i-1}, S^{1:N}.Y^{1:N}).$$

  - Decode $\hat{U}_i'$ as 0 for the block corresponding to the component V_N of vector $V^{1:N}$ in q staircases.
  - Now we have at least L positions of the MDS codeword that are recovered. Now the erasure decoding of the MDS code recovers all $N-1$ positions of the codeword.
  - Hence all $\hat{U}_i'$s corresponding to all polar blocks in the column are estimated in all q staircases. This enables the continuation of SC decoding for the polar blocks to estimate $\hat{U}_i'$s corresponding to the next column.
- Decoding non-full-height columns on the right side.
  - The U_is corresponding to these columns are transmitted as is by the encoder.
  - Hence $\hat{U}_i' = U_i - E_i' = U_i$ for all these columns in all q staircases.
- Now $\hat{U}^{1:N} = \hat{U}'^{1:N} + E'^{1:N}$ for each polar block.
- $\hat{X}^{1:N} = \hat{U}^{1:N} G_N$ for each block.

Theorem 1.
The probability of error for the above staircase scheme is $O(Nqk2^{-N^{\beta}})$ for $\beta < 0.5$.

Proof:
We decode $U'^{1:N}$ corresponding to all the polar blocks. The error occurs if and only if there is an error in decoding some good-bit-channel ($\mathcal{L}_{X|Y}$) of any polar block. If U_is for good bit-channels are recovered properly, then other U_is are recovered either by MDS erasure decoding in a full-height column or by the knowledge of U_is at the receiver corresponding to the non-full-height columns. Let the error event be $\mathcal{E}$.

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 $i \in \mathcal{L}_{X|Y}$ for all polar blocks. Let all the polar blocks in all of the q staircases be indexed as $b = 1, 2..., Nqk$. Let $\mathcal{E}_{ib}$ be the error event corresponding to an error in the ith bit-channel of block b. If bit-channel $i \in \mathcal{L}_{X|Y}$ of the polar block b lies in a full-height column, then the error event $\mathcal{E}_{ib}$ becomes as follows.

$$\begin{aligned}\mathcal{E}_{ib} = \{(U'^{1:N}, Y^{1:N}, S^{1:N}) \text{ of block } b :\\ \Phi_i(U'^{1:i-1}, S^{1:N}.Y^{1:N}) \neq U_i'\}.\end{aligned}$$

Note that $\mathcal{E}_{ib}$ will be the null event, if the block b has bit-channel i that lies in a non-full-height column. Clearly, $\mathcal{E}_g = \cup_{b\in\{1:Nqk\}} \cup_{i\in\{1:N\}} \mathcal{E}_{ib}$. Note that error event $\mathcal{E}$ will imply at least one of the $\mathcal{E}_{ib}$s. So we should have the following.

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

Now the probability of error $P(\mathcal{E})$ is upper bounded as follows.

$$\begin{aligned}P(\mathcal{E}) \leq P(\mathcal{E}_g) &= P(\cup_{b\in\{1:Nqk\}} \cup_{i\in\{1:N\}} \mathcal{E}_{ib})\\ &\overset{(a)}{\leq} \sum_{b\in\{1:Nqk\}} \sum_{i\in\mathcal{L}_{X|Y}} P(\mathcal{E}_{ib}).\end{aligned}$$

The identity (a) follows from the union bound. So, we need to bound $P(\mathcal{E}_{ib})$ for $i \in \mathcal{L}_{X|Y}$ for all polar blocks.

Now we evaluate the conditional probability of error of bit-channel $i \in \mathcal{L}_{X|Y}$ for the block b given the random vectors $(U'^{1:N}, Y^{1:N}, S^{1:N})$ corresponding to the block b.

$$\begin{aligned}&P(\mathcal{E}_{ib}|U'^{1:N} = u'^{1:N}, S^{1:N} = s^{1:N})\\ &= P(\Phi_i(U'^{1:i-1}, S^{1:N}.Y^{1:N}) \neq U_i'|\\ &\qquad\qquad U'^{1:N} = u'^{1:N}, S^{1:N} = s^{1:N})\\ &= P(\Phi_i(u'^{1:i-1}, s^{1:N}.Y^{1:N}) \neq u_i'|\\ &\qquad\qquad U'^{1:N} = u'^{1:N}, S^{1:N} = s^{1:N})\\ &\overset{(a)}{=} \sum_{y^{1:N}} \Pi_{i\in[1:N]} p(y_i|(u'^{1:N}G_N)_i)\\ &\qquad\qquad \mathbb{1}(\Phi_i(u'^{1:i-1}, y^{1:N}) \neq u_i')\\ &\overset{(b)}{\leq} Z_i\\ &= 2^{-N^{\beta}}.\end{aligned} \tag{1}$$

The identity (a) follows from Lemma 2. Identity (b) follows from Arikan's [1] symmetric channel polar coding construction where it was proved that the conditional probability of error of a bit-channel given that any particular word is transmitted over the channel is always the same irrespective of the word that is transmitted. Also, all of those conditional probabilities of errors are upper bounded by the Bhattacharyya parameter of the bit-channel. This is essentially stated as Fact 1. Now the actual probability of error of bit-channel i for the block b satisfies

$$\begin{aligned}P(\mathcal{E}_{ib}) &= \sum_{u'^{1:N}, s^{1:N}} P(U'^{1:N} = u'^{1:N}, S^{1:N} = s^{1:N})\\ &\qquad\qquad P(\mathcal{E}_{ib}|U'^{1:N} = u'^{1:N}, S^{1:N} = s^{1:N})\\ &\leq \sum_{u'^{1:N}, s^{1:N}} P(U'^{1:N} = u'^{1:N}, S^{1:N} = s^{1:N})2^{-N^{\beta}}\\ &= 2^{-N^{\beta}}.\end{aligned}$$

Therefore,

$$P(\mathcal{E}) \leq \sum_{b\in\{1:Nqk\}} \sum_{i\in\mathcal{L}_{X|Y}} P(\mathcal{E}_{ib})$$
$$\leq O(Nqk2^{-N^{\beta'}}).$$

Hence the proof of Theorem 1. □

B. Coding with non-uniform source

If the conditional distributions in $\mathcal{C}$ are of symmetric type, then any rate greater than $\max_{p(y|x)\in\mathcal{C}} H(\tilde{X}|Y)$ can still be achieved using the staircase method irrespective of the source distribution $P_X(x)$. Here the random variable pair $(\tilde{X}, Y)$ is distributed as $P_{\tilde{X}}(x)p(y|x)$ and $P_{\tilde{X}}(x) = 0.5$. The subtle idea is to implement the same code construction as if the source is uniformly distributed. We use the bit-channels $\mathcal{L}_{\tilde{X}|Y}$ in the code construction irrespective of the source distribution. The conditional probability of error of the bit-channel $i \in \mathcal{L}_{\tilde{X}|Y}$ is the same given any source sequence due to the symmetric channel property of the conditional distribution $p(y|x)$. Hence the average probability of error for the bit-channel $i \in \mathcal{L}_{\tilde{X}|Y}$ does not depend on the source distribution. This can be noticed from equation (1) in the proof of Theorem 1. Therefore the probability of error will still be $O(Nkq2^{-N^{\beta}})$ for $\beta < 0.5$.

C. Encoding and Decoding Complexity

The encoding complexity consists of decomposing the vector of length $N-1$ in GF(2^q) into a RS codeword and the corresponding error vector. We proposed to use the shift register circuit with adders and multipliers to get a systematic RS codeword for executing the decomposition. This takes $O(L(N-L)) = O(N^2)$ multiplications and additions in GF(2^q). Addition and multiplication over this field take q and $q^{\log_2 3}$ binary operations, respectively. Hence the bit operations sum upto $O(N^2 q^{\log_2 3})$ for each full-height column. Applying polar transform for each polar block takes $O(N \log_2 N)$ bit-operations.

Decoding complexity consists of applying the polar transform ($S^{1:N} = E'^{1:N} G_N$) for all polar blocks, SC decoding of all the polar blocks and also the erasure decoding of the RS codes of length $N-1$ over GF(2^q) for each full-height column. Applying polar transform for each polar block takes $O(N\log_2 N)$ bit-operations. The SC decoding of a polar block takes $O(N \log_2 N)$ real operations. The erasure decoding of the RS codes can be done in O($N(\log_2(N))^2$) symbol operations [4]. Addition and multiplication over this field take q and $q^{\log_2 3}$ binary operations respectively. Hence the bit operations sum to $O(N(\log_2(N))^2 q^{\log_2 3})$ for each full-height column.

D. Pros and Cons

Pros:
The upside of the scheme is that it can be designed for a class $\mathcal{C}$ with infinite cardinality as well. In particular, the block length does not increase with cardinality of the class $\mathcal{C}$. On the other hand, a code designed with rate R supports any arbitrary source with any side information whose conditional distribution given source $p(y|x)$ is of symmetric channel type whenever $R > H(\tilde{X}|Y)$. Here the random variable pair $(\tilde{X}, Y)$ is distributed as $P_{\tilde{X}}(x)p(y|x)$ and $P_{\tilde{X}}(x) = 0.5$.

Cons:
The downside is that the class $\mathcal{C}$ has to contain only conditional distributions of symmetric channel type. Also, for the non-uniform source with distribution $P_X(x)$, the staircase construction does not support all the rates greater than $\max_{p(y|x)\in\mathcal{C}} H(X|Y)$ where the random variable pair (X, Y) is distributed as $P_X(x)p(y|x)$.

V. Scheme based on combining bit-channels

In this scheme, we assume the class $\mathcal{C}$ contains a finite number of elements. Let $|\mathcal{C}|$ be s. The bit-channel sets $\mathcal{L}_{X|Y}$ and $\mathcal{H}_{X|Y}$ will not be the same for each $p(y|x) \in \mathcal{C}$. The obvious approach is to share $U^{(\cap_{i\in\mathcal{C}}\mathcal{L}_{X|Y})^c}$, so that decoder can reliably decode the other bits corresponding to bit-channels in $(\cap_{i\in\mathcal{C}}\mathcal{L}_{X|Y})$ by SC decoding. The scheme based on bit-channel combining is a recursive procedure of combining polar blocks that increases the fraction of bit-channels $\cap_{p(y|x)\in\mathcal{C}}\mathcal{L}_{X|Y}$ with respect to the updated polar block length. The fraction of bit-channels $\cap_{p(y|x)\in\mathcal{C}}\mathcal{L}_{X|Y}$ with respect to the updated polar block length in the recursive procedure can get arbitrarily close to $1 - \max_{p(y|x)\in\mathcal{C}} H(X|Y)$ when N is sufficiently large. Hence, this gives the compression algorithm that can achieve any rate greater than $\max_{p(y|x)\in\mathcal{C}} H(X|Y)$. Hassani and Uranke [4] essentially did this for a symmetric source in the context of universal channel coding. We need to validate that such a recursive method can be used for a non-uniform memoryless source setting as well. So, this method is straightforward to use in this source coding setting in view of the original scheme [4] proposed in the context of universal channel coding.

We need Proposition 1 to validate this method for an arbitrary discrete memoryless source (which may be non-uniform) with the arbitrary class $\mathcal{C}$ (which may contain non-symmetric $p(y|x)$) of finite cardinality.

Lemma 3. *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.

The Lemma 3 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 $\mathcal{X} = \{0,1\}$, while 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.

We now validate the method with an arbitrary memoryless source while recalling the idea of this method proposed in [4]. Let $\mathcal{C} = \{p_1(y|x), p_2(y|x), ..., p_s(y|x)\}$. The first step is to increase the fraction of bit-channels $\mathcal{L}_{X|Y_1} \cap \mathcal{L}_{X|Y_2}$ with respect to the updated block length. To do this, first consider the two independent polar blocks $U^{1:N} = X^{1:N}G_N$ and $U'^{1:N} = X'^{1:N}G_N$, where $Y^{1:N}$ and $Y'^{1:N}$ are the correlated side information vectors corresponding to the two blocks, respectively. Then combine the bit-channels $\mathcal{L}_{X|Y_1} \cap \mathcal{H}_{X|Y_2}$ of the first block with bit-channels $\mathcal{L}_{X|Y_2} \cap \mathcal{H}_{X|Y_1}$ in the order. Suppose the bit-channel $i \in \mathcal{L}_{X|Y_1} \cap \mathcal{H}_{X|Y_2}$ with input U_i and output $U^{1:i-1}Y^{1:N}$ from the first polar block is combined with bit-channel $j \in \mathcal{L}_{X|Y_2} \cap \mathcal{H}_{X|Y_1}$ with input U'_j and output $U'^{1:j-1}Y'^{1:N}$ from the second polar block. One of the two new bit-channels produced by this combining has the input $U_i + U'_j$ and the output $U^{1:i-1}U'^{1:j-1}Y^{1:N}Y'^{1:N}$; the other bit-channel produced has the input U'_j and the output $U_i + U'_j, U^{1:i-1}U'^{1:j-1}Y^{1:N}Y'^{1:N}$. By Proposition 1, the second bit-channel produced by the combining has the Bhattacharyya parameter

$$Z(U'_i|U_i + U'_j, U^{1:i-1}U'^{1:j-1}Y^{1:N}Y'^{1:N})$$
$$= Z(U_i|U^{1:i-1}Y^{1:N})Z(U'_j|U^{1:j-1}Y^{1:N})) \overset{(a)}{\leq} O(2^{-N^{\beta}}).$$

where $\beta < 0.5$. The identity (a) is true because either the Bhattacharyya parameter $Z(U_i|U^{1:i-1}Y^{1:N})$ is $2^{-N^{\beta}}$ if the conditional distribution is $p_1(y|x)$ or the Bhattacharyya parameter $Z(U'_j|U'^{1:j-1}Y^{1:N})$ is $2^{-N^{\beta}}$ if the conditional distribution is $p_2(y|x)$. So we have $G = \min\{|\mathcal{L}_{X|Y_2} \cap \mathcal{H}_{X|Y_1}|, |\mathcal{L}_{X|Y_1} \cap \mathcal{H}_{X|Y_2}\}|$ new bit-channels that come into the category of $\mathcal{L}_{X|Y_1} \cap \mathcal{L}_{X|Y_2}$ in the updated polar block of length $2N$. We use a bold font from now on to denote the bit-channels in the updated polar block to disitnguish them from the bit-channels of the original polar block. Now the fraction of the updated bit-channels $\mathcal{L}_{\mathbf{X|Y_1}} \cap \mathcal{L}_{\mathbf{X|Y_2}}$ with respect to the updated block length becomes as follows.

$$\frac{2(\mathcal{L}_{X|Y_1} \cap \mathcal{L}_{X|Y_2}) + G}{2N}.$$

The procedure can be done recursively. In stage t of the recursive procedure, we take two polar blocks obtained in stage $t-1$ and perform the same bit-channel combinings that were mentioned in the first step. After t recursions, the fraction of the updated $\mathcal{L}_{\mathbf{X|Y_1}} \cap \mathcal{L}_{\mathbf{X|Y_2}}$ with respect to the updated block length becomes as follows.

$$\frac{2^t(\mathcal{L}_{X|Y_1} \cap \mathcal{L}_{X|Y_2}) + (2^t - 1)G}{2^tN}.$$

This will increase and become closer to $|\mathcal{L}_{X|Y_1} \cap \mathcal{L}_{X|Y_2}| + G = \min\{|\mathcal{L}_{X|Y_1}|, |\mathcal{L}_{X|Y_2}|\}$ per block length N as t grows. Now let the bit-channels $\mathcal{L}_{\mathbf{X|Y_1}} \cap \mathcal{L}_{\mathbf{X|Y_2}}$ in the updated polar block be $\mathcal{L}_{\mathbf{12}}$ and repeat the same recursive procedure to increase the bit-channels $\mathcal{L}_{12} \cap \mathcal{L}_{X|Y_3}$. We continue the recursive procedure until we finish all $p(y|x) \in \mathcal{C}$. Hence by this method, one can increase the cardinality of bit-channels $\cap_{p(y|x)\in\mathcal{C}}\mathcal{L}_{\mathbf{X|Y}}$ per block length N that can get arbitrarily close to $\min_{p(y|x)\in\mathcal{C}} |\mathcal{L}_{X|Y}|$. The details for this method are given in [4]. Our main requirement is to show the validity of this method with an arbitrary source, which followed from Proposition 1. The scheme supports any non-uniform source with an arbitrary class $\mathcal{C}$ of finite cardinality. But the block length can become unbounded as the cardinality of the class $\mathcal{C}$ grows, in contrast to the staircase scheme.

VI. Conclusion

We defined the problem of source coding with side information at the receiver whose correlation is unknown to the encoder. We studied two coding strategies based on polar codes for this problem. The code designed by the staircase scheme with rate R supports any source with any side information whose conditional distribution given source $p(y|x)$ is of symmetric channel type whenever $R > H(\tilde{X}|Y)$. Here the random variable pair $(\tilde{X}, Y)$ is distributed as $P_{\tilde{X}}(x)p(y|x)$ and $P_{\tilde{X}}(x) = 0.5$. A naive Slepian-Wolf code derived using the method [11] requires the computation of a high dimensional systematic parity-check matrix $(q[(N-L)(1+N(k-1)) + N(N-1)] \times N^2qk)$ for the universal channel code. We avoid the computation of such a high dimensional parity-check matrix and its use in our staircase code construction. The second scheme is based on the technique of universalization using bit-channel combining. Using this method, we can design a code for a non-uniform source with arbitrary $\mathcal{C}$ of finite cardinality. An open problem is to find a stronger coding strategy where a code designed for an arbitrary source X at rate R can support any correlated side information Y whenever $R > H(X|Y)$.

VII. 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] E. Arıkan, “Source polarization,” *Proc. IEEE Int. Symp. Inf. Theory,* Austin, TX, 2010, pp. 899-903.

[3] E. Arikan,"Polar coding for the Slepian-Wolf problem based on monotone chain rules," *Proc. IEEE Int. Symp. Inf. Theory,* Cambridge, MA, 2012, pp. 566-570.

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

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

[6] 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.

[7] S. B. Korada and R. Urbanke, "Polar codes for Slepian-Wolf, Wyner-Ziv, and Gelfand-Pinsker," *Proc. IEEE Information Theory Workshop (ITW)*, Cairo, 2010, pp. 1-5.

[8] K. Nagarjuna and P. H. Siegel, "Universal polar coding for asymmetric channels," *Proc. IEEE Information Theory Workshop (ITW)*, Guangzhou, Nov. 2018, pp. 1-5.

[9] K. Nagarjuna and P. H. Siegel, *Universal polar coding for asymmetric channels*, *http://cmrr-star.ucsd.edu/static/pubs/asymmetric_upc.pdf*

[10] R. Roth, *Introduction to Coding Theory*. Cambridge University Press, 2006

[11] L. Wang and Y. Kim, "Linear code duality between channel coding and Slepian-Wolf coding," *Proc. 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, 2015, pp. 147-152.

Appendix A

Proof of Lemma 3:

$$\begin{aligned} Z(X|Y) &= 2\sum_{y\in\mathcal{Y}} \sqrt{P_{X,Y}(0,y)P_{X,Y}(1,y)} \\ &= -1 + \sum_{y\in\mathcal{Y}} \big[\sum_{x\in\mathcal{X}} \sqrt{P_{X,Y}(x,y)}\big]^2 \\ &\overset{(a)}{\geq} -1 + \sum_{y\in\mathcal{Y}}\sum_{j\in\mathcal{J}} Q(j)\big[\sum_{x\in\mathcal{X}} \sqrt{P^j_{X,Y}(x,y)}\big]^2 \qquad (2) \\ &= \sum_{j\in\mathcal{J}} Q(j)(-1 + \sum_{y\in\mathcal{Y}} \big[\sum_{x\in\mathcal{X}} \sqrt{P^j_{X,Y}(x,y)}\big]^2) \\ &= \sum_{j\in\mathcal{J}} Q(j) Z^j(X|Y). \end{aligned}$$

Inequality (a) follows from Minkowsky's inequality

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

which holds when $r < 1$ and a_{jk} is non-negative. 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.

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

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

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

be the two joint distributions on random variable triplet (U_1,Y_1,Y_2). Then,

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

Identity (a) is true because Y_1 is independent of U_2 and 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 3 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 $Z(U_1|Y_1Y_2) \geq \max\{Z(X_2|Y_2), Z(X_1|Y_1)\}$. Then,

$$\begin{aligned} &Z(U_2|Y_1Y_2U_1) \\ &= 2\sum_{y_1y_2u_1\in\mathcal{Y}_1\times\mathcal{Y}_2\times\mathcal{X}} P_{U_1Y_1Y_2}(u_1y_1y_2) \\ &\qquad \sqrt{P_{U_2|U_1Y_1Y_2}(0|u_1y_1y_2)P_{U_2|U_1Y_1Y_2}(1|u_1y_1y_2)} \\ &= 2\sum_{y_1y_2u_1\in\mathcal{Y}_1\times\mathcal{Y}_2\times\mathcal{X}} \frac{P_{U_1Y_1Y_2}(u_1y_1y_2)}{P_{U_1|Y_1Y_2}(u_1|y_1y_2)} \\ &\qquad \big[P_{X_1|Y_1}(u_1|y_1)P_{X_2|Y_2}(0|y_2) \\ &\qquad\quad P_{X_1|Y_1}(u_1+1|y_1)P_{X_2|Y_2}(1|y_2)\big]^{0.5} \\ &= 2\sum_{y_1y_2u_1\in\mathcal{Y}_1\times\mathcal{Y}_2\times\mathcal{X}} P_{Y_1Y_2}(y_1y_2) \\ &\qquad \big[P_{X_1|Y_1}(u_1|y_1)P_{X_1|Y_1}(u_1+1|y_1) \\ &\qquad\quad P_{X_2|Y_2}(0|y_2)P_{X_2|Y_2}(1|y_2)\big]^{0.5} \\ &= 2\sum_{u_1\in\mathcal{X}}\sum_{y_1\in\mathcal{Y}_1}\sum_{y_2\in\mathcal{Y}_2} P_{Y_1}(y_1)P_{Y_2}(y_2) \\ &\qquad \big[P_{X_1|Y_1}(u_1|y_1)P_{X_2|Y_2}(0|y_2) \\ &\qquad\quad P_{X_1|Y_1}(u_1+1|y_1)P_{X_2|Y_2}(1|y_2)\big]^{0.5} \\ &= Z(X_1|Y_1)Z(X_2|Y_2). \end{aligned}$$

□