On the Polarization of Rényi Entropy

Mengfan Zheng

Based on joint work with Ling Liu and Cong Ling

Dept. of Electrical and Electronic Engineering
Imperial College London

m.zheng@imperial.ac.uk

8 May, 2019

Motivation

- Shannon entropy/Mutual information
  - measure information in the average sense
  - work well in communication theory
  - insufficient in some other areas such as cryptography

- Rényi entropy: more general, widely adopted in cryptography, etc.

- Polarization/polar codes: powerful tool, well-studied under Shannon's information measures

- Polarization of Rényi entropy not well understood yet

Outline

Notations

- $(X, Y) \sim P_{X,Y}$
- $[N]$: index set $\{1, 2, ..., N\}$.
- Vectors: $\mathbf{X}$ or $X^{a:b} \triangleq \{X^a, X^{a+1}, ..., X^b\}$ where $a \leq b$.
- $X^{\mathcal{A}}$ ($\mathcal{A} \subset [N]$): the subvector $\{X^i : i \in \mathcal{A}\}$ of $X^{1:N}$.
- $\mathbf{G}_N = \mathbf{B}_N \mathbf{F}^{\otimes n}$: the generator matrix of polar codes, where $N = 2^n$, $\mathbf{B}_N$ is the bit-reversal matrix, and $\mathbf{F} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$.

Shannon Entropy

- (Shannon) Entropy:

$$H(X) = \mathbb{E}_P \log \frac{1}{P(X)} = -\sum_{x \in \mathcal{X}} P(x) \log P(x)$$

- Joint entropy:

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log P(y, x)$$

- Conditional entropy:

$$H(Y|X) = \sum_{x \in \mathcal{X}} P(x) H(Y|X = x) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log P(y|x)$$

- Chain rule:

$$H(X, Y) = H(X) + H(Y|X)$$

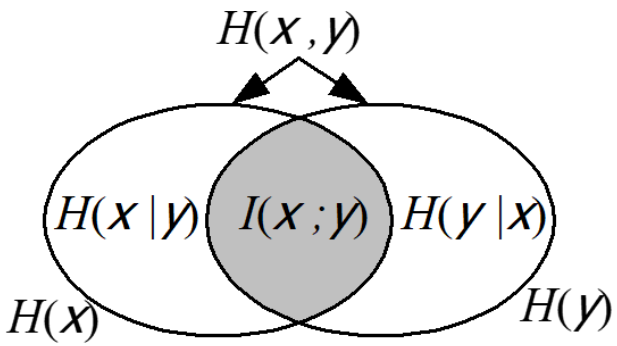
Relative Entropy and Mutual Information

- The relative entropy or Kullback–Leibler distance between two probability mass functions $P(x)$ and $Q(x)$:

$$D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = \mathbb{E}_P \log \frac{P(x)}{Q(x)}$$

- Mutual information: the average information that Y gives about X

$$I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} = D(P(x, y)||P(x)P(y))$$



From Shannon to Rényi

Definition (Rényi Entropy [Rényi'61])

The Rényi entropy of a random variable $X \in \mathcal{X}$ of order α is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} P_X(x)^\alpha. \tag{1}$$

As $\alpha \to 1$, the Rényi entropy reduces to the Shannon entropy.

Three other special cases of the Rényi entropy:

- *Max-entropy*: $H_0(X) = \log |\mathcal{X}|$
- *Min-entropy*: $H_\infty(X) = \min_i(-\log p_i) = -\log \max_i p_i$
- *Collision entropy*: $H_2(X) = -\log \sum_{i=1}^n p_i^2 = -\log P(X = Y)$

Rényi Entropy

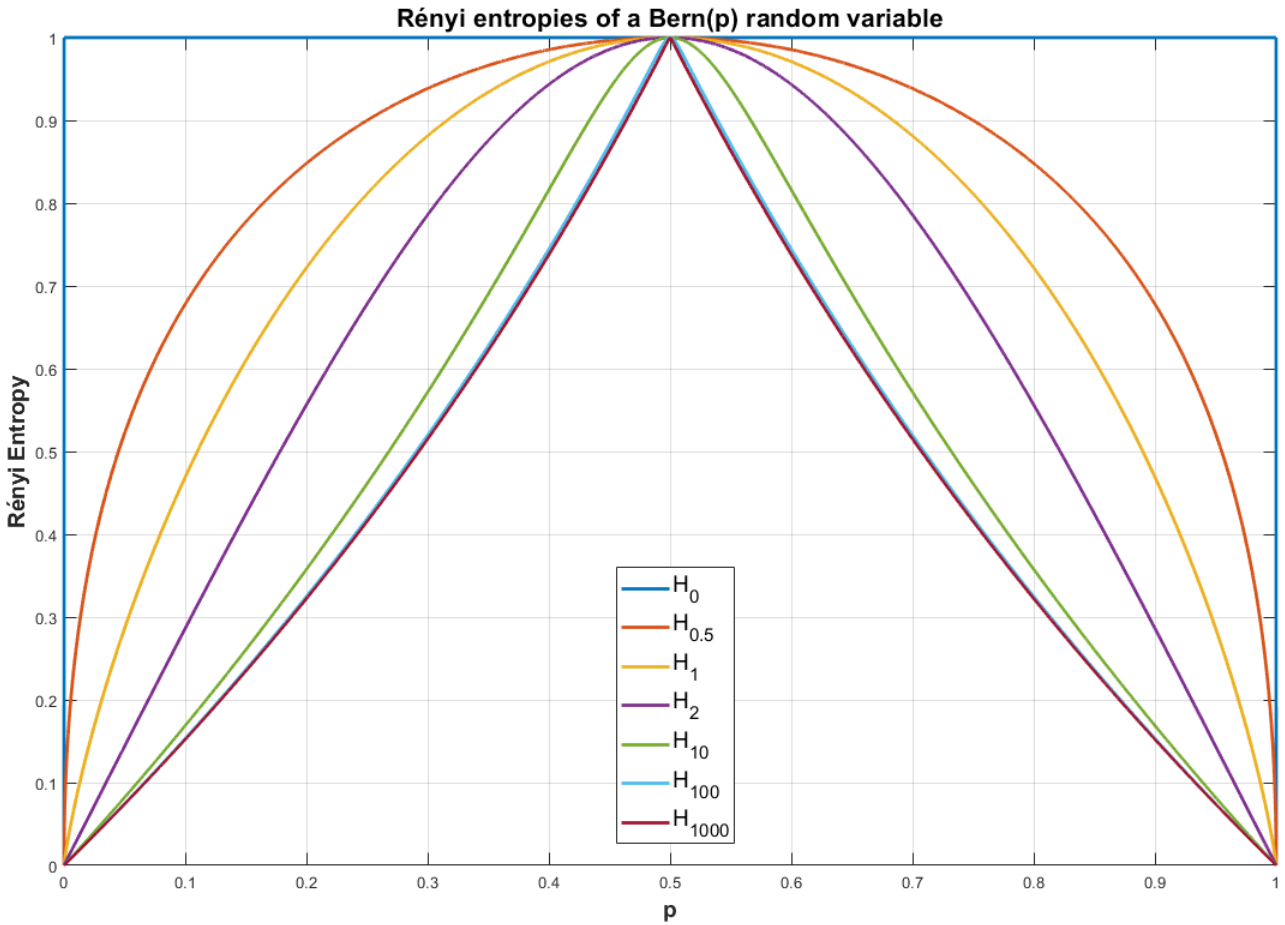


Figure: Rényi entropies of a $Bern(p)$ random variable.

Rényi Divergence

Definition (Rényi divergence [Rényi'61])

The Rényi divergence of order α of P from another distribution Q on $\mathcal{X}$ is defined as

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \sum_{x \in \mathcal{X}} P(x)^\alpha Q(x)^{1-\alpha}. \tag{2}$$

Also, as $\alpha \to 1$, the Rényi divergence reduces to the Kullback–Leibler divergence.

Conditional Rényi Entropy

Unlike the conditional Shannon entropy, there is no generally accepted definition of the conditional Rényi entropy yet.

Definition (Conditional Rényi Entropy [Jizba-Arimitsu'04])

The conditional Rényi entropy of order α of X given Y is defined as

$$H_\alpha(X|Y) = \frac{1}{1-\alpha} \log \frac{\sum_{\{x,y\}\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)^\alpha}{\sum_{y\in\mathcal{Y}} P_Y(y)^\alpha}. \tag{3}$$

This type of Rényi conditional entropy satisfies the chain rule:

$$H_\alpha(X|Y) + H_\alpha(Y) = H_\alpha(X,Y). \tag{4}$$

Conditional Rényi Entropy (Cont.)

Definition (Conditional Rényi Entropy [Cachin'97])

The conditional Rényi Entropy of order α of X given Y is defined as

$$H'_\alpha(X|Y) = \sum_{y\in\mathcal{Y}} P_Y(y) H_\alpha(X|y). \tag{5}$$

Conditional Rényi Entropy (Cont.)

Definition (Conditional Rényi Entropy [Arimoto'77])

The conditional Rényi Entropy of order α of X given Y is defined as

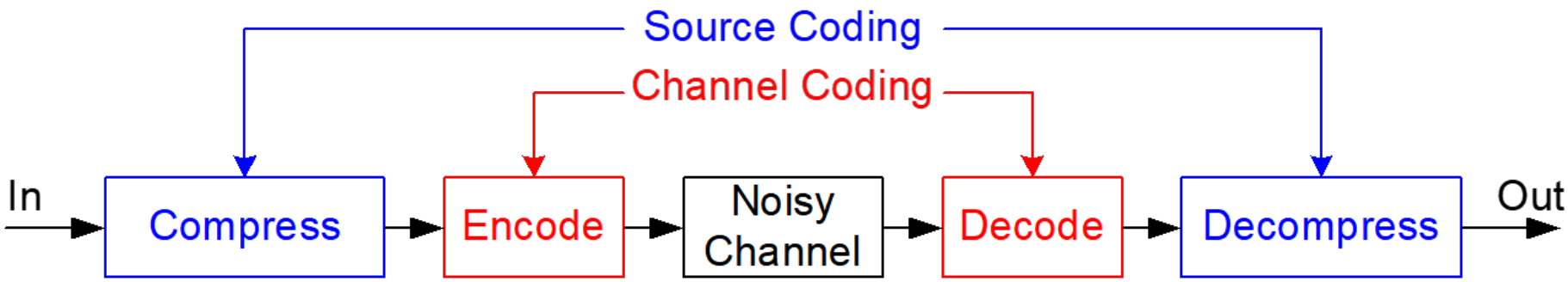
$$H_\alpha^A(X|Y) = \frac{\alpha}{1-\alpha} \log \sum_{y\in\mathcal{Y}} P_Y(y) \Big[\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)^\alpha \Big]^{\frac{1}{\alpha}} \tag{6}$$

Definition (Conditional Rényi Entropy [Hayashi'11])

The conditional Rényi Entropy of order α of X given Y is defined as

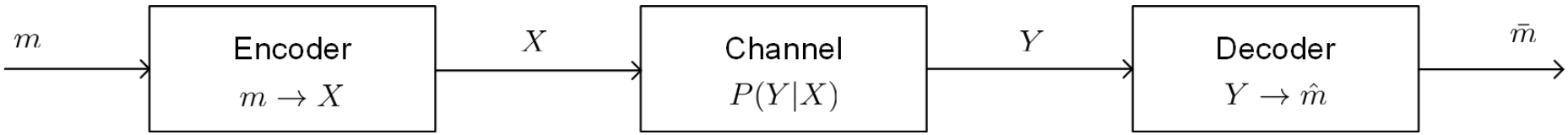
$$H_\alpha^H(X|Y) = \frac{1}{1-\alpha} \log \sum_{y\in\mathcal{Y}} P_Y(y) \sum_{x\in\mathcal{X}} P_{X|Y}(x|y)^\alpha \tag{7}$$

Model of Digital Communication



- Source Coding
  - Compresses the data to remove redundancy
- Channel Coding
  - Adds redundancy/structure to protect against channel errors

The Channel Coding Problem



- $m \in \mathcal{M} = \{1, 2, ..., M\}$
- Input $X \in \mathcal{X}$, output $Y \in \mathcal{Y}$
- Memoryless: $P(y^n|x^{1:n}, y^{1:n-1}) = P(y^n|x^n)$
- DMC = Discrete Memoryless Channel

Channel Capacity

- Capacity of a DMC channel: $C = \max_{P_x} I(X;Y)$

  – Mutual information (not entropy itself) is what could be transmitted through the channel
  – Maximum is over all possible input distributions P_x
  – $\exists$ only one maximum since $I(X;Y)$ is concave in $P(x)$ for fixed $P(y|x)$
  – We want to find the P_x that maximizes $I(X;Y)$
  – Limits on C:

$$0 \leq C \leq \min(H(X), H(Y)) \leq \min(\log|\mathcal{X}|, \log|\mathcal{Y}|)$$

- Capacity for n uses of channel:

$$C^{(n)} = \frac{1}{n} \max_{P_{X^{1:n}}} I(X^{1:n}; Y^{1:n})$$

Channel Coding

- Assume Discrete Memoryless Channel (DMC) with known $P_{Y|X}$.

- An (M, n)-code is
  - A fixed set of M codewords $\mathbf{x}(w) \in \mathcal{X}^n$ for $w = 1 : M$
  - A deterministic decoder $g(\mathbf{y}) \in 1 : M$

- The rate of an (M, n)-code: $R = (\log_2 M)/n$ bits/transmission

- Error probability: $\lambda_w = P(g(\mathbf{y}(w)) \neq w) = \sum_{\mathbf{y} \in \mathcal{Y}^n} P(\mathbf{y}|\mathbf{x}(w)) \delta_{g(\mathbf{y}) \neq w}$
  - Maximum error probability:

$$\lambda^{(n)} = \max_{1 \leq w \leq M} \lambda_w$$

  - Average error probability:

$$P_e^{(n)} = \frac{1}{M} \sum_{w=1}^{M} \lambda_w$$

Shannon's ideas

- Channel coding theorem: the basic theorem of information theory
  - Proved in his original 1948 paper

- How do you correct all errors?

- Shannon's ideas
  - Allow an arbitrarily small but nonzero error probability
  - Use the channel many times in succession, so that the law of large numbers comes into effect
  - Consider a randomly chosen code and show the expected average error probability is small
  - Use the idea of typical sequences
  - Show this means $\exists$ at least one code with small max error prob

- Sadly it does not tell you how to construct the code

About Polar Codes

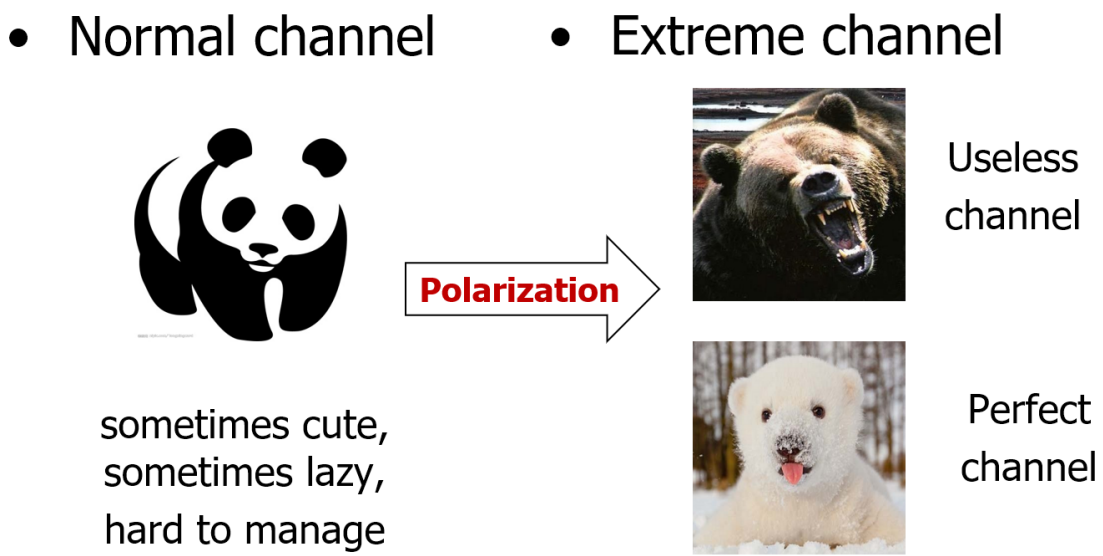
- First low-complexity scheme which provably achieves the capacity of binary-input memoryless symmetric channels.
- Encoding complexity $O(N \log N)$
- Successive decoding complexity $O(N \log N)$
- Probability of error $\approx 2^{-\sqrt{N}}$
- Not only good for channel coding, works equally well for source coding and more complicated scenarios.
- Main idea: **channel/source polarization**

What is Polarization

- Among all channels, there are two classes which are easy to communicate optimally
  - The perfect channels: the output Y determines the input X
  - The useless channels: Y is independent of X

- Polarization is a technique to convert noisy channels to a mixture of extreme channels

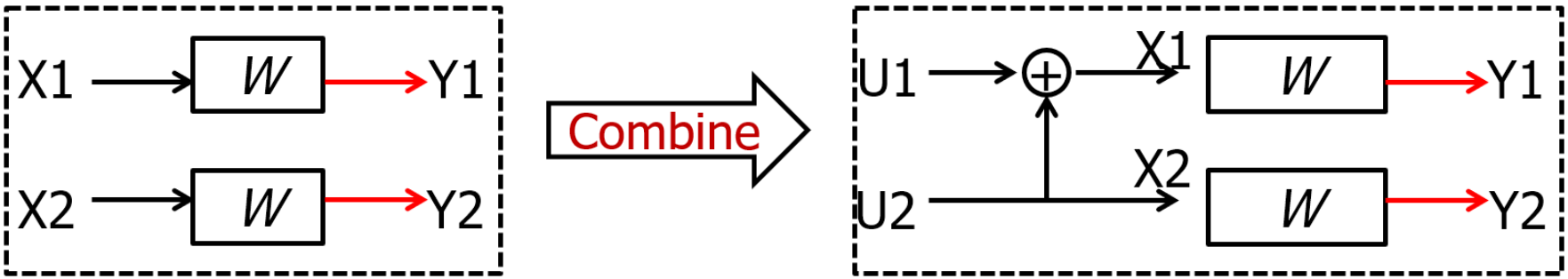
- The process is information-conserving



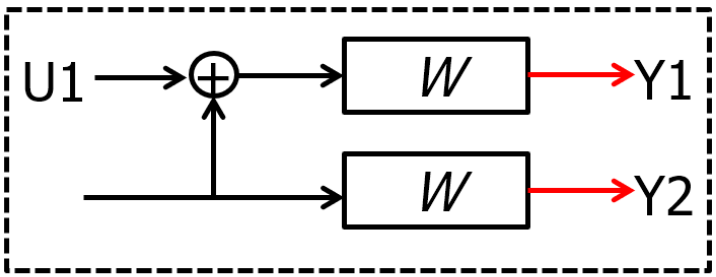
Channel Combining and Splitting

Basic operation ($N = 2$)

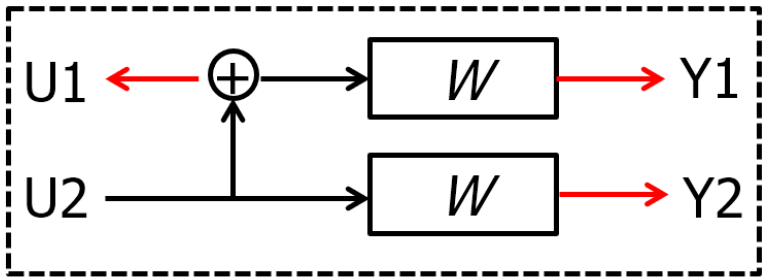
- Channel combining:



- Channel splitting:



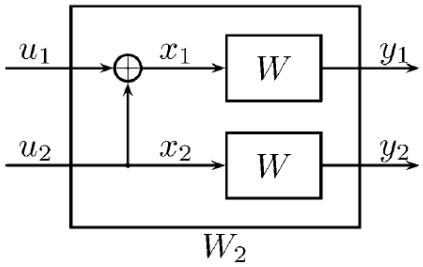
W^-: U1→(Y1, Y2)

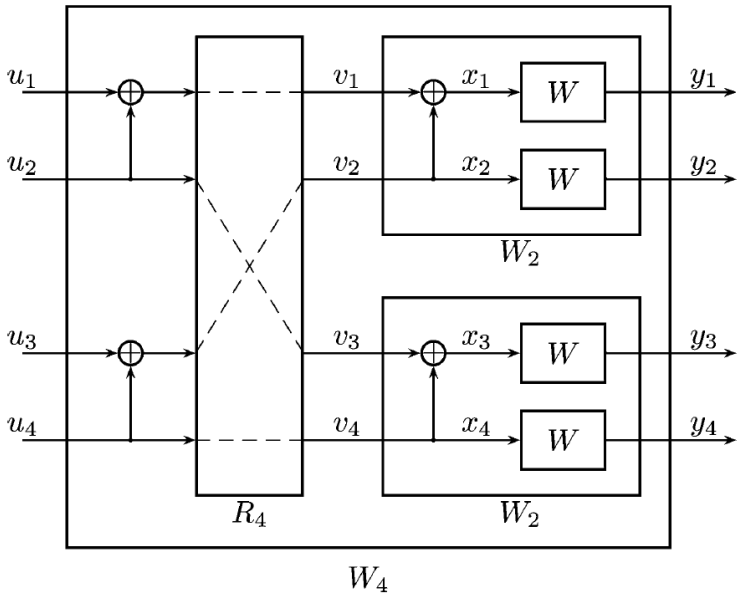


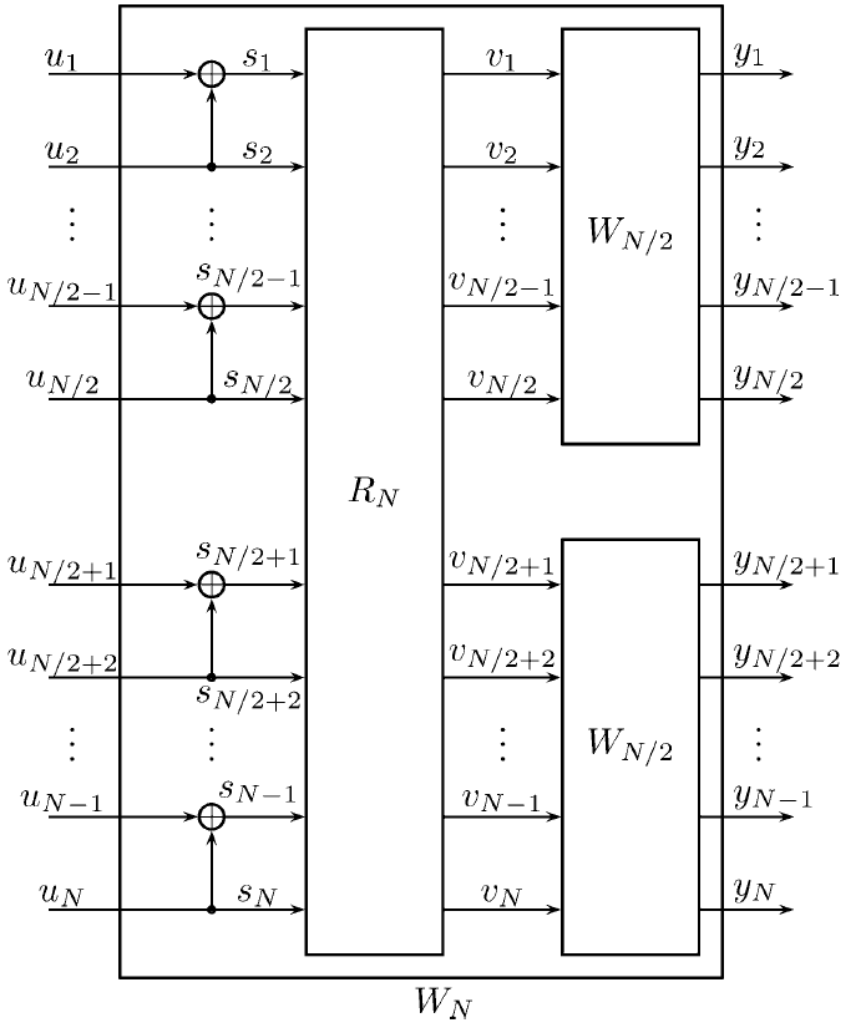
W^+: U2→(Y1, Y2, U1)

Polar Transformation

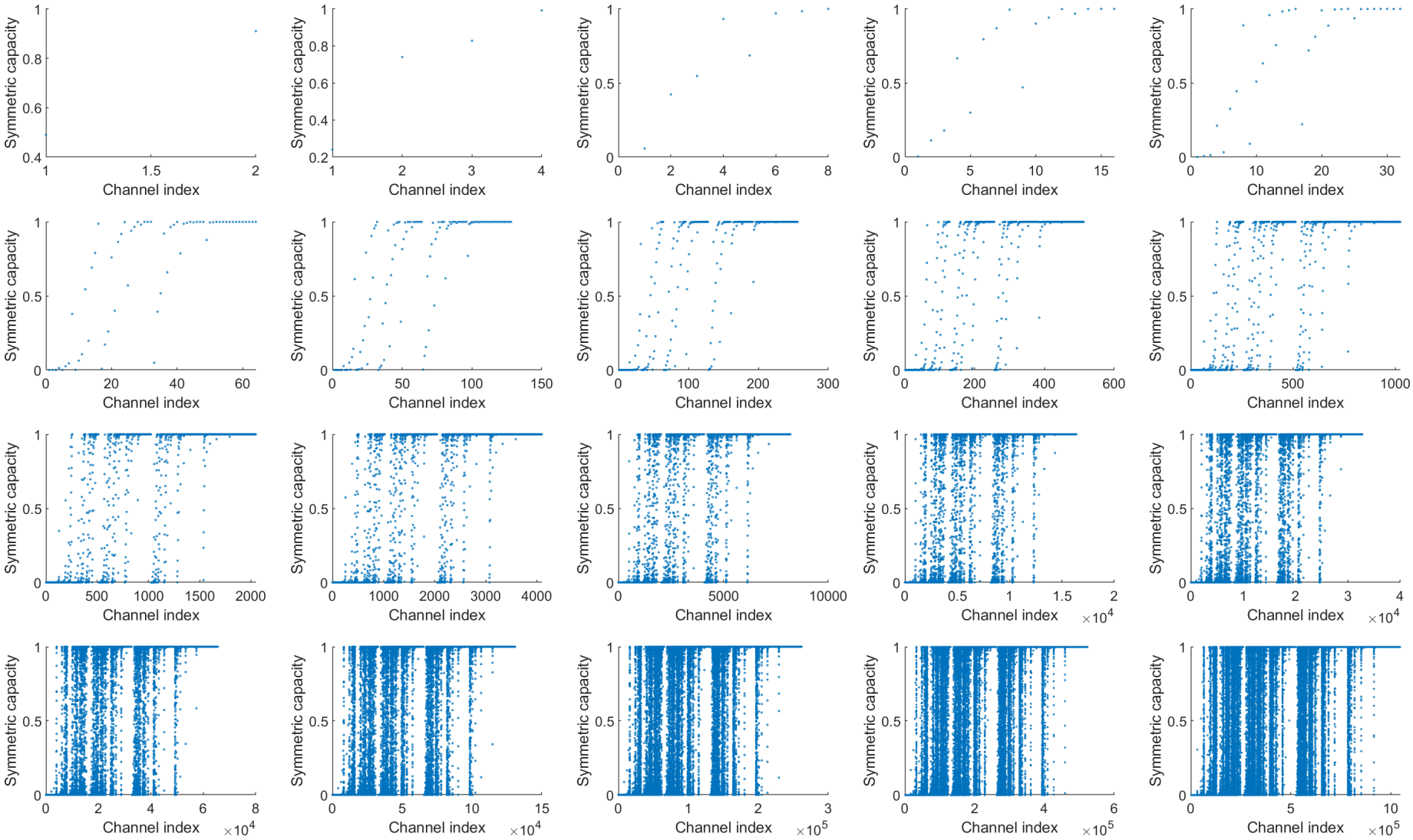
Recursive polar transformations [Arıkan'09]:



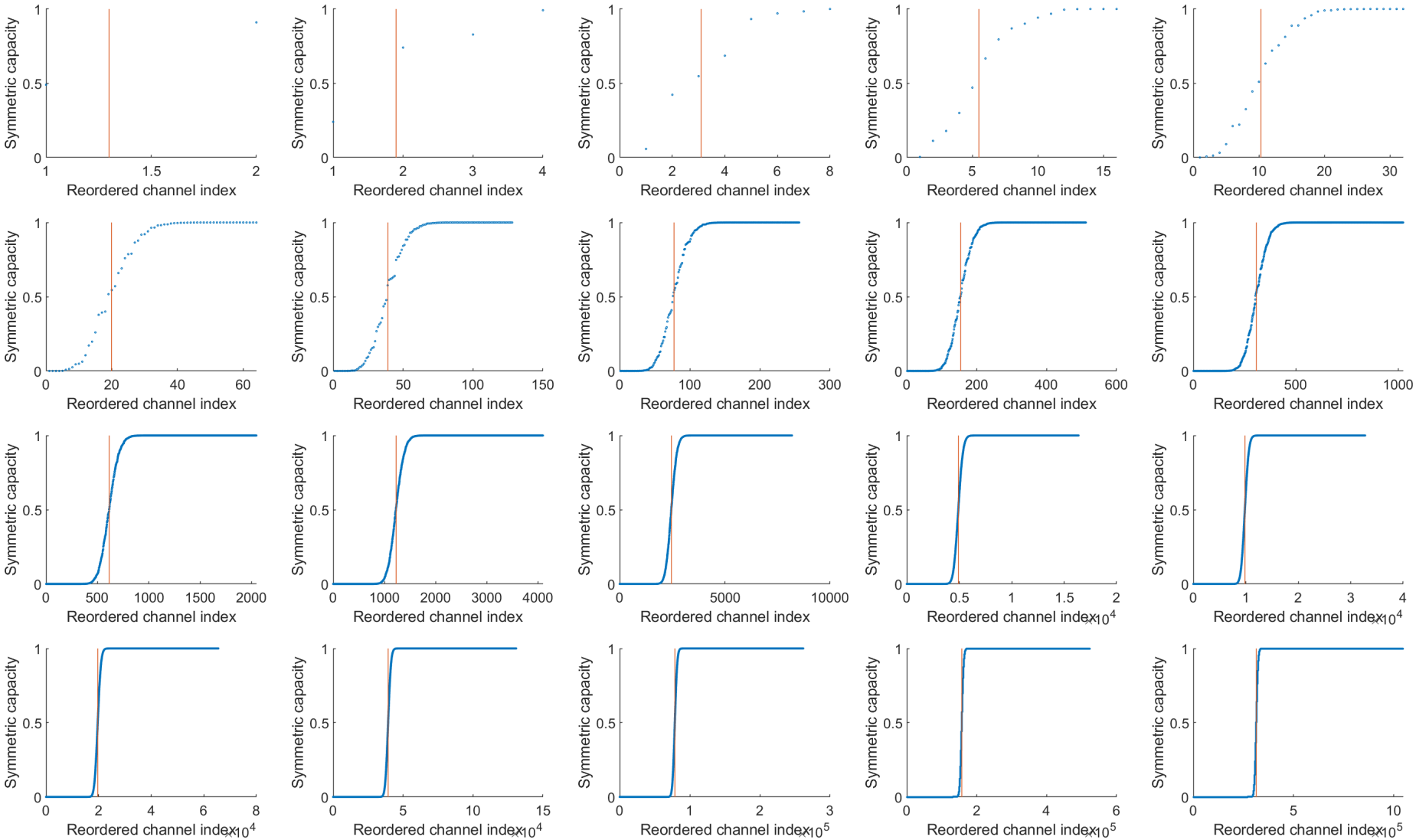




Polarization of BEC(0.3), n=1:20



Polarization of BEC(0.3), n=1:20 (Cont.)



Polarization Theory

Theorem (Channel polarization [Arıkan'09])

For any B-DMC W, the channels $\{W_N^{(i)}\}$ polarize in the sense that, for any fixed $\delta \in (0,1)$, as $N \to \infty$ through the power of 2, the fraction of indices $i \in [N]$ for which $I(W_N^{(i)}) \in (1-\delta, 1]$ goes to $I(W)$, and the fraction with $I(W_N^{(i)}) \in [0, \delta)$ goes to $1 - I(W)$.

Polar Codes

- $X^{1:N}$: N consecutive channel inputs to a B-DMC $W(Y|X)$.
- Let $U^{1:N} = X^{1:N}\mathbf{G}_N$, where $N = 2^n$, $\mathbf{G}_N = \mathbf{B}_N\mathbf{F}^{\otimes n}$.
- For $\delta_N = 2^{-N^\beta}$ with $\beta \in (0, 1/2)$, define

$$\mathcal{I} = \{i \in [N] : H(U^i|Y^{1:N}, U^{1:i-1}) \leq \delta_N\}. \tag{8}$$

- Assign $\{u^i\}_{i\in\mathcal{I}}$ with information bits, and $\{u^i\}_{i\in\mathcal{I}^C}$ with frozen bits. Then compute $x^{1:N} = u^{1:N}\mathbf{G}_N$.
- Upon receiving $y^{1:N}$, the receiver uses a successive cancellation (SC) decoder to decode:

$$\bar{u}^i = \begin{cases} u^i, & i \in \mathcal{I}^C \\ \arg\max_{u\in\{0,1\}} P_{U^i|Y^{1:N}U^{1:i-1}}(u|y^{1:N}, u^{1:i-1}), & i \in \mathcal{I} \end{cases}. \tag{9}$$

Existing Results

• In [Alsan-Telatar'14] it is shown that the following chain rule inequality holds for the polar transformation for $\alpha \leq 1$,

$$H_\alpha^*(U_1 U_2|Y_1 Y_2) \geq H_\alpha^*(U_1|Y_1 Y_2) + H_\alpha^*(U_2|Y_1 Y_2 U_1),$$

whenever U_1, U_2 are i.i.d. uniform on $\mathbb{F}_2$.

• The inequality holds with equality if and only if the channel W is perfect, or the channel W is completely noisy, or $\alpha = 1$.

• Note that $H_\alpha^*(X|Y)$ in [Alsan-Telatar'14] is defined as

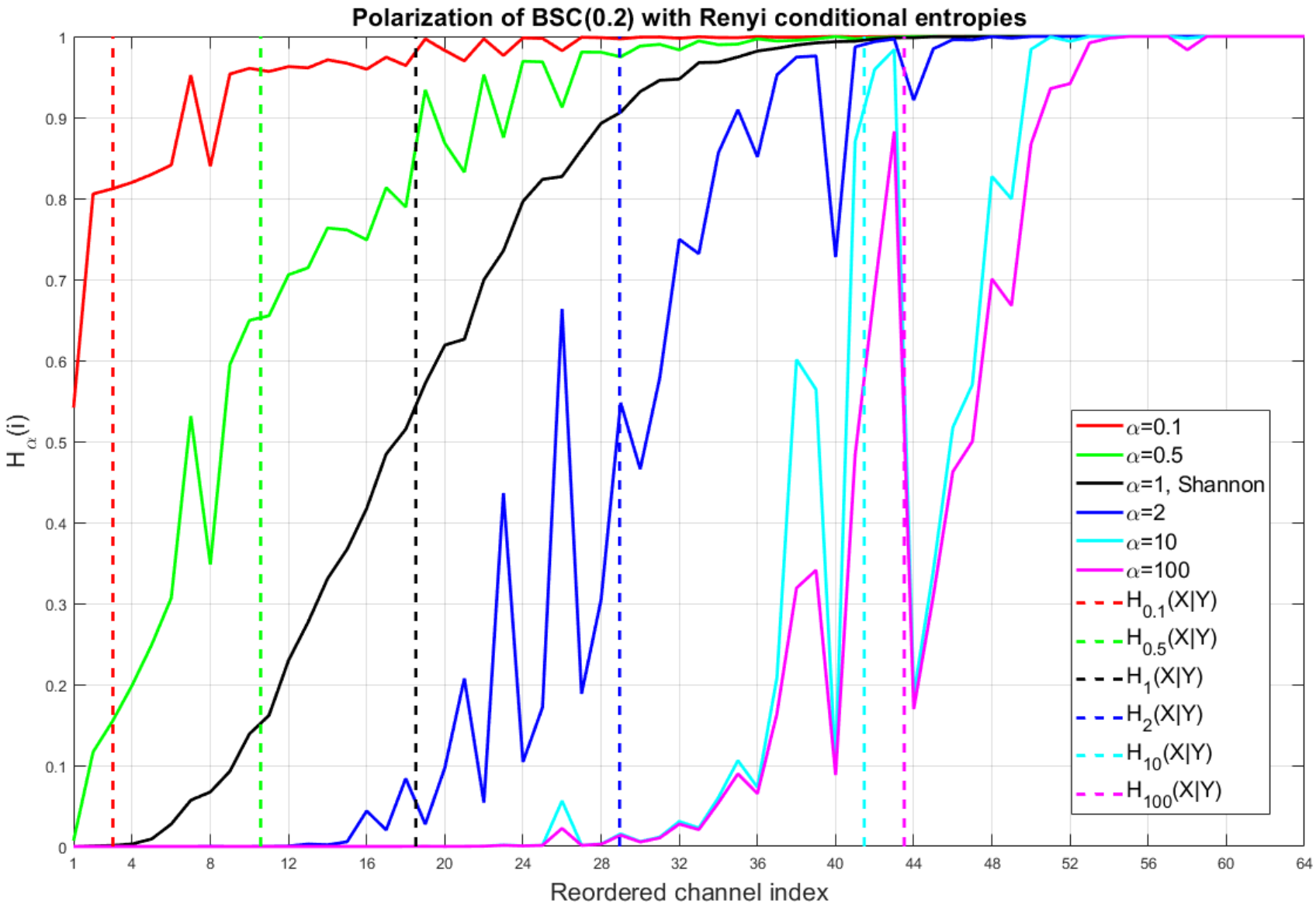
$$H_\alpha^*(X|Y) = H_\alpha(X) + \frac{\alpha}{1-\alpha} \log \sum_{y \in \mathcal{Y}} \Big[\sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}(y|x)^\alpha \Big]^{\frac{1}{\alpha}}.$$

Polarization of Conditional Rényi Entropy (Cont.)

Theorem (Polarization of Conditional Rényi Entropy)

For any B-DMC $P_{Y|X}$ (or any discrete memoryless source $(X, Y) \sim P_{X,Y}$ over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y}$ an arbitrary countable set) and any $\alpha \geq 0$, as $N \to \infty$ through the power of 2, the fraction of indices $i \in [N] \triangleq \{1, 2..., N\}$ with $H_N(i) \in (1 - \delta, 1]$ goes to $H_\alpha(X|Y)$, and the fraction with $H_N(i) \in [0, \delta)$ goes to $1 - H_\alpha(X|Y)$.

Polarization of Conditional Rényi Entropy (Cont.)



Proof

Lemma (Basic polar transformation)

For $\alpha \geq 0$, we have

$$H_\alpha(U_2|Y_1Y_2U_1) \leq \min\{H_\alpha(X_1|Y_1), H_\alpha(X_2|Y_2)\}. \tag{10}$$

$$H_\alpha(U_1|Y_1Y_2) \geq \max\{H_\alpha(X_1|Y_1), H_\alpha(X_2|Y_2)\}, \tag{11}$$

$$H_\alpha(U_1U_2|Y_1Y_2) = H_\alpha(U_1|Y_1Y_2) + H_\alpha(U_2|Y_1Y_2U_1). \tag{12}$$

Proof (Cont.)

Define a random walk $\{B_n; n \geq 0\}$ in the infinite binary tree

- The random walk starts at the root node with $B_0 = (0, 1)$, and moves to one of the two child nodes in the next level with equal probability at each integer time.
- If $B_n = (n, i)$, B_{n+1} equals $(n+1, 2i-1)$ or $(n+1, 2i)$ with probability $1/2$ each.

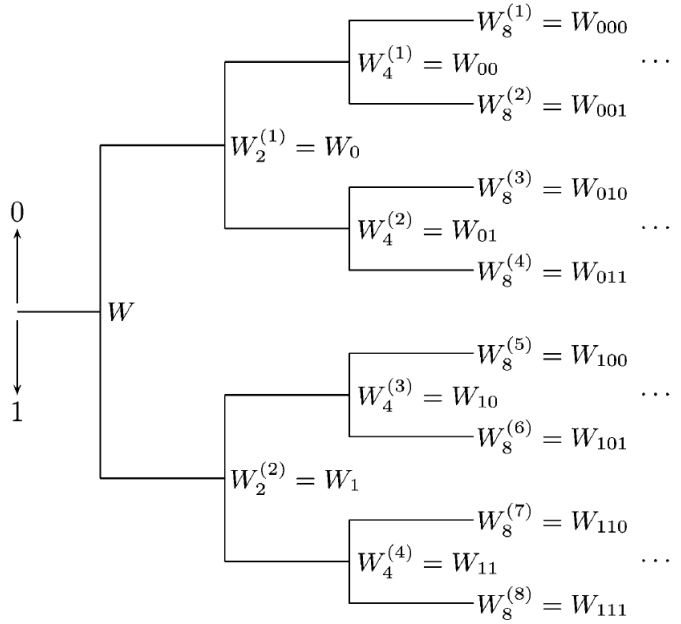


Figure: The tree process for the recursive channel construction [Arıkan'09].

Proof (Cont.)

• Denote $H(0,1) = H_\alpha(X|Y)$ and $H(n,i) = H_\alpha(U^i|Y^{1:2^n}, U^{1:i-1})$ for $n \geq 1$, $i = [2^n]$, and define a random process $\{H_n; n \geq 0\}$ with $H_n = H(B_n)$.

• The random process $\{H_n; n \geq 0\}$ is a martingale due to the chain rule equality of (12), i.e.,

$$E[H_{n+1}|B_0, B_1, ..., B_n] = H_n.$$

• Since $\{H_n; n \geq 0\}$ is a uniformly integrable martingale, it converges a.e. to an RV H_∞ such that $E[|H_n - H_\infty|] = 0$. Then we have

$$E[|H_n - H_{n+1}|] \to 0. \tag{13}$$

Proof (Cont.)

- Note that

$$
\begin{aligned}
& E[|H_n - H_{n+1}|] \\
= & \frac{1}{2}\Big(E[|H_\alpha(U^i|Y^{1:2^n}, U^{1:i-1}) - H_\alpha(U^{2i-1}|Y^{1:2^{n+1}}, U^{1:2i-2})|] \\
& + E[|H_\alpha(U^i|Y^{1:2^n}, U^{1:i-1}) - H_\alpha(U^{2i}|Y^{1:2^{n+1}}, U^{1:2i-1})|]\Big).
\end{aligned}
$$

Thus, (13) forces the inequalities in (10) and (11) to hold with equalities.

- We can then show that in this case, $H_\alpha(U^i|Y^{1:2^n}, U^{1:i-1})$ tends to either 0 or 1 as $n \to \infty$.

- The convergence result together with the chain rule equality of (12) imply that the fraction of $\{i : H(n,i) \in (1-\delta, 1]\}$ goes to $H_\alpha(X|Y)$ as $n \to \infty$.

Discussion

- Under Rényi entropies of different orders, the same synthetic sub-channel may exhibit opposite extremal states.

- Intuitively, if a synthetic sub-channel is totally deterministic or uniform, its Rényi entropies of different orders should be the same.

- A paradox?

An Example

- Let $|\mathcal{Y}| = 2^N \triangleq M$.

- Consider such a $P_{X,Y}$: $\frac{1}{L}$ of probability pairs $\big(P_{X,Y}(0,y), P_{X,Y}(1,y)\big)$ are completely deterministic with accumulated probability $\frac{1}{N}$.

- Without loss of generality, assume $P_{X,Y}(0,y_i) = \frac{L}{NM}$ and $P_{X,Y}(1,y_i) = 0$ for $i \in \mathcal{A}$, where $\mathcal{A} \subset [2^N]$ is the set of deterministic pairs.

- The rest $\frac{L-1}{L}$ fraction of probability pairs are completely uniform, i.e., $P_{X,Y}(0,y_i) = P_{X,Y}(1,y_i) = \frac{(N-1)L}{2NM(L-1)}$ for $i \in \mathcal{A}^C$.

An Example (Cont.)

- Then

$$
\begin{aligned}
H_\alpha(X|Y) &= \frac{1}{1-\alpha}\log\frac{\sum_{i\in\mathcal{A}}\left(\frac{L}{NM}\right)^\alpha + 2^{1-\alpha}\sum_{i\in\mathcal{A}^c}\left(\frac{(N-1)L}{NM(L-1)}\right)^\alpha}{\sum_{i\in\mathcal{A}}\left(\frac{L}{NM}\right)^\alpha + \sum_{i\in\mathcal{A}^c}\left(\frac{(N-1)L}{NM(L-1)}\right)^\alpha} \\
&= \frac{1}{1-\alpha}\log\frac{\frac{M}{L}\left(\frac{L}{NM}\right)^\alpha + 2^{1-\alpha}\frac{M(L-1)}{L}\left(\frac{(N-1)L}{NM(L-1)}\right)^\alpha}{\frac{M}{L}\left(\frac{L}{NM}\right)^\alpha + \frac{M(L-1)}{L}\left(\frac{(N-1)L}{NM(L-1)}\right)^\alpha} \\
&= \frac{1}{1-\alpha}\log\frac{(L-1)^{\alpha-1} + 2^{1-\alpha}(N-1)^\alpha}{(L-1)^{\alpha-1} + (N-1)^\alpha}. \qquad (14)
\end{aligned}
$$

An Example (Cont.)

- For a considered $\alpha = \alpha_0 > 1$, let

$$L - 1 = 2^{-1}(N-1)^{\frac{\alpha_0 - 0.5/\alpha_0}{\alpha_0 - 1}}.$$

- Then we have

$$H_{\alpha_0}(X|Y) = \frac{1}{1-\alpha_0} \log \frac{2^{1-\alpha_0}\left[(N-1)^{\alpha_0 - \frac{0.5}{\alpha_0}} + (N-1)^{\alpha_0}\right]}{2^{1-\alpha_0}(N-1)^{\alpha_0 - \frac{0.5}{\alpha_0}} + (N-1)^{\alpha_0}}.$$

- It is clear that as $N \to \infty, \frac{(N-1)^{\alpha_0}}{(N-1)^{\alpha_0 - \frac{0.5}{\alpha_0}}} = (N-1)^{0.5/\alpha_0} \to \infty$.
- Thus $H_{\alpha_0}(X|Y) \to 1$.

An Example (Cont.)

- Now let $\alpha' = \alpha_0 + 1$. From (14) we have

$$H_{\alpha'}(X|Y) = \frac{1}{-\alpha_0} \log \frac{2^{-\alpha_0}(N-1)^{\frac{\alpha_0^2-0.5}{\alpha_0-1}} + 2^{-\alpha_0}(N-1)^{\alpha_0+1}}{2^{-\alpha_0}(N-1)^{\frac{\alpha_0^2-0.5}{\alpha_0-1}} + (N-1)^{\alpha_0+1}}.$$

- In this case, as $N \to \infty$, $\frac{(N-1)^{\alpha_0+1}}{(N-1)^{\frac{\alpha_0^2-0.5}{\alpha_0-1}}} = (N-1)^{\frac{-0.5}{\alpha_0^2-1}} \to 0$.

- Thus $H_{\alpha'}(X|Y) \to 0$.

An Example (Cont.)

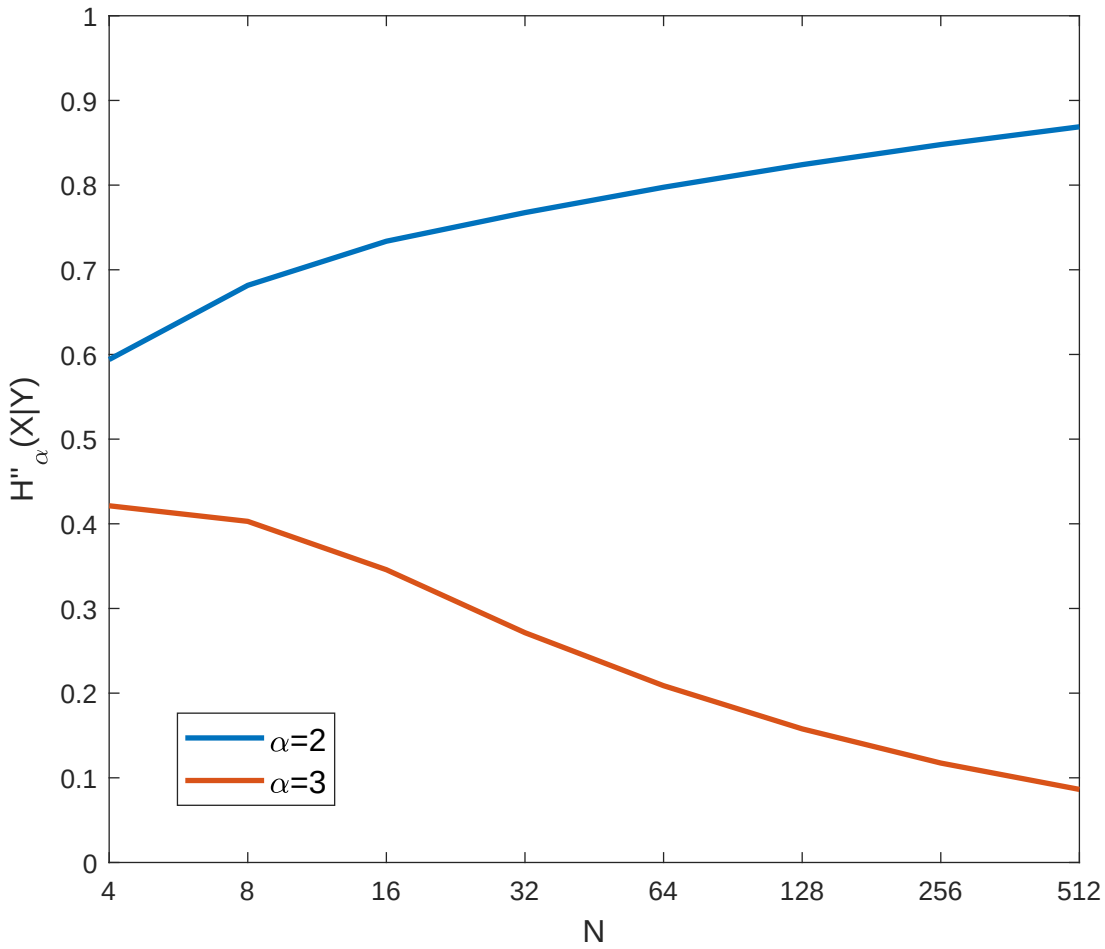


Figure: Conditional Rényi entropies of different α.

Secure Coding Schemes

- Weak security scheme (requires $\frac{1}{N} I(Z^{1:N}; M) \to 0$):

  – Secret information bit positions chosen from

$$\mathcal{L}_{X|Y}^{(N)} \setminus \mathcal{L}_{X|Z}^{(N)}$$

  where

- Strong security scheme (requires $I(Z^{1:N}; M) \to 0$):

  – Secret information bit positions chosen from

$$\mathcal{L}_{X|Y}^{(N)} \cap \mathcal{H}_{X|Z}^{(N)}$$

$$\mathcal{L}_{X|Y}^{(N)} = \{i \in [N] : H(U^i|Y^{1:N}, U^{1:i-1}) \leq \delta_N\}$$
$$\mathcal{H}_{X|Z}^{(N)} = \{i \in [N] : H(U^i|Z^{1:N}, U^{1:i-1}) \geq 1 - \delta_N\}$$
$$\mathcal{L}_{X|Z}^{(N)} = \{i \in [N] : H(U^i|Z^{1:N}, U^{1:i-1}) \leq \delta_N\}$$

Secure Coding Schemes (Cont.)

- A more general definition of secrecy: ϵ-secrecy with respect to $H_\alpha(\cdot|\cdot)$ [Iwamoto-Shikata'13]

$$H_\alpha(M) - H_\alpha(M|C) \leq \epsilon$$

- Can we design a polar code that satisfies this condition?

Randomness Extraction

- Source polarization:

Let $X^{1:N} = [X^1, ..., X^N]$ be i.i.d. $Bern(p)$, $N = 2^n$, and $U^{1:N} = X^{1:N}\mathbf{G}_N$. Then, for any $\epsilon \in (0, 1)$,

$$\frac{1}{N}|\{j \in [N] : H(U^j|U^{1:j-1}) \geq 1 - \epsilon\}| \to H(p)$$

- Using polarization to do randomness extraction was discussed in [Abbe'11].

- Secret-key generation is an application [Chou et al'15].

Open Problems

- No chain rule equality or inequality
- The polarization rate of Rényi entropy?
- How to determine the overall Rényi divergence?
- What about other types of conditional Rényi entropy?

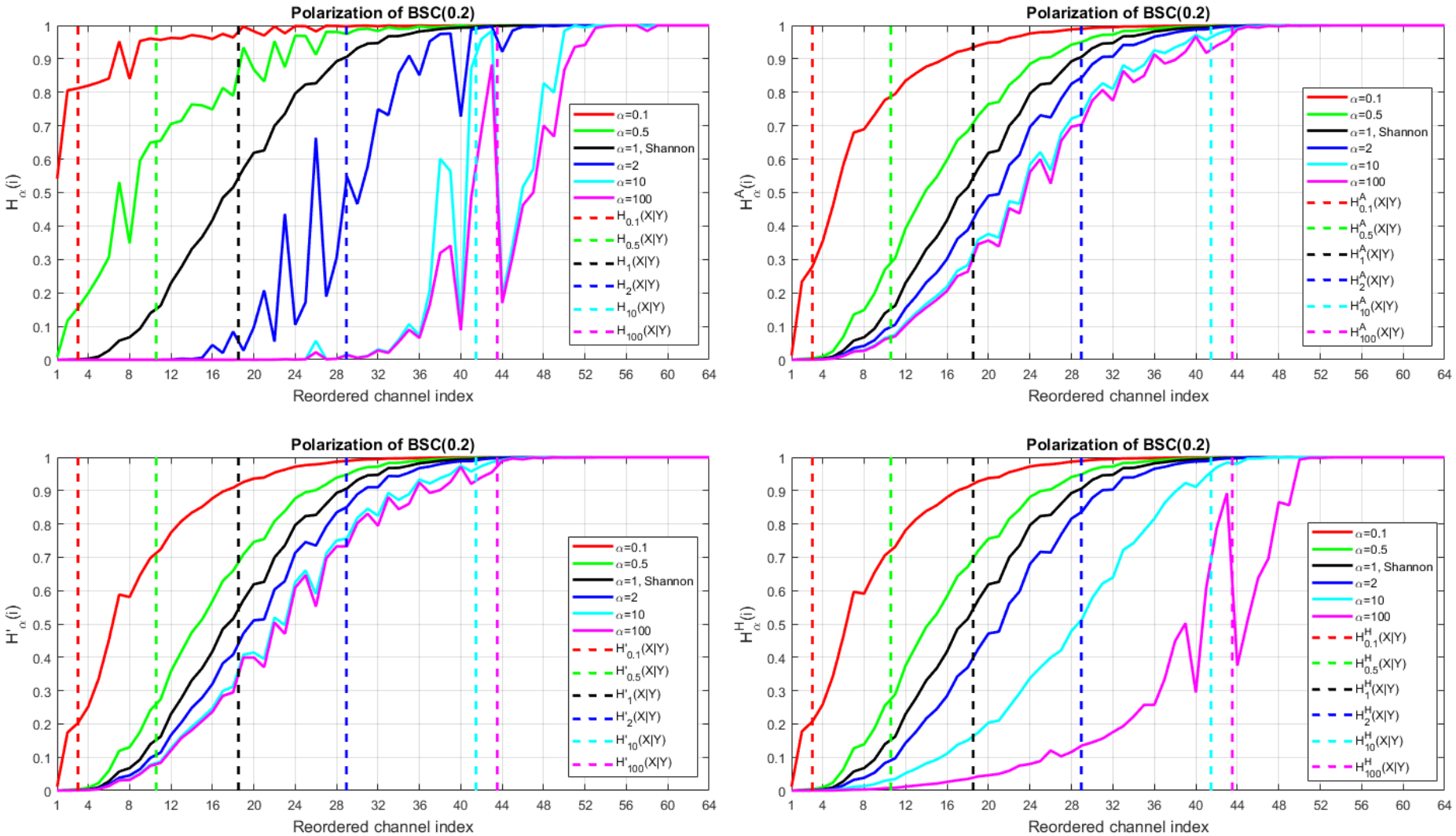


Figure: Polarization of different types of conditional Rényi entropies.

References

A. Rényi (1961)
On measures of entropy and information
Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics.

P. Jizba and T. Arimitsu (2004)
The world according to Rényi: thermodynamics of multifractal systems
Annals of Physics, 312(1), 17-59.

S. Arimoto (1977)
Information measures and capacity of order α for discrete memoryless channels
Topics in information theory.

M. Hayashi (2011)
Exponential decreasing rate of leaked information in universal random privacy amplification
IEEE Transactions on Information Theory, 57(6), 3989-4001.

C. Cachin (1997)
Entropy measures and unconditional security in cryptography
Doctoral dissertation, ETH Zurich.

References (Cont.)

E. Arıkan (2009)
Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels
IEEE Transactions on Information Theory, 55(7), 3051-3073.

M. Alsan and E. Telatar (2014)
Channel Polarization: Polarization Improves E_0
IEEE Transactions on Information Theory, 60(5), 2714-2719.

M. Iwamoto and J. Shikata, (2013)
Information theoretic security for encryption based on conditional Rényi entropies
IEEE Transactions on Information Theory, 60(5), 2714-2719.

E. Abbe (2011)
Randomness and dependencies extraction via polarization
2011 Information Theory and Applications Workshop, (pp. 1-7).

R. A. Chou, M. R. Bloch and E. Abbe (2015)
Polar coding for secret-key generation
IEEE Transactions on Information Theory, 61(11), 6213-6237.

Thank you!