Channel Polarization and Polar Codes

by

Mine Alsan

Supervisor: Prof. Emre Telatar

Technical Report
Information Theory Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne

February 2012

Abstract

Polar coding is a new technique introduced by Arıkan [1] based on a phenomenon called channel polarization. Polar codes are appealing as an error correction method since they are proved to achieve the symmetric capacity of any B-DMC using low complexity encoders and decoders, and their block error probability is shown to decrease exponentially in the square root of the block length. In fact, two basic channel transformations lie at the heart of channel polarization. The recursive applications of these transformations result in channel polarization which refers to the fact that the channels synthesized in these transformations become in the limit either almost perfect or completely noisy. This is shown by analyzing channel parameters such as the symmetric capacity and the Bhattacharya parameter. An important characteristic of polar codes is that they are channel specific codes. For that particular reason, the channel over which we communicate information should be considered during the code design. However, the binary erasure channel and the binary symmetric channel stand out as extremal channels among all B-DMCs in terms of the evolution of the channel symmetric capacity and the Bhattacharyya parameter under the basic channel transformations. In this work, we generalize this extremality result to a more general parameter $E_0(\rho, W)$, defined by Gallager [2] as part of the random coding exponent, for a B-DMC W with uniform input distribution. We show that the binary erasure channel and the binary symmetric channel are also extremal with respect to the evolution of the parameter $E_0(\rho, W)$ under the basic channel transformations. Then, we conjecture an inequality between $E_0(\rho, W)$, $E_0(\rho, W^-)$, and $E_0(\rho, W^+)$. In the process, we note that the function $E_0(\rho, W)/\rho$ is interpreted as a general measure of information using Rényi's entropy functions. Moreover, we discuss an application on the compound capacity of polar codes under successive cancellation decoding where the extremality of the binary erasure channel is used to derive a lower bound. We also provide a discussion on the compound capacity of linear codes which includes polar codes as sub-class. We show that while linear codes achieving the compound capacity of symmetric channels exist, the existing results on the compound capacity of polar codes decoded using a successive cancellation decoder shows that, in general, the compound capacity of symmetric channels is not achieved by polar codes. In addition, independently of channel polarization, we undertake a study of another channel property: the random coding exponent in parametric form. Note that this parametric description is a function of $E_0(\rho, W)$ and the rate $R(\rho, W) = \partial E_0(\rho, W)/\partial\rho$. We extend the binary erasure channel and the binary symmetric channel extremality result in [3] in terms of $E_0(\rho, W)$ and the rate $R(\rho, W)$ to the case where we have different ρ values, i.e., $E_0(\rho_1, W)$ and $R(\rho_2, W)$.

Contents

List of Figures

Chapter 1

Introduction

1.1 Channel Polarization and Polar Codes

Consider the communication scenario described in Figure 1.1 over a binary discrete memoryless channel (B-DMC) W with binary input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probabilities $P(y \mid x)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Given the binary data sequence $U_1^n = U_1 \dots U_n$, the encoder generates the binary input sequence $\mathcal{E}(U_1^n) = X_1^n = X_1 \dots X_n$ applying a one-to-one transformation $G : \{0,1\}^n \to \{0,1\}^n$ and transmits this sequence over W. The decoder receives the channel output sequence $Y_1^n = Y_1 \dots Y_n$, and produces an estimate $\hat{U}_1 \dots \hat{U}_n = \mathcal{D}(Y_1^n)$ of the initial data. Further, assume that the input data sequence is i.i.d from a binary uniform distribution. Consequently, the channel input sequence is also i.i.d with binary uniform distribution, and $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d pairs of random variables.

Figure 1.1: Communication over a B-DMC W.

We want to design the channel encoder and decoder such that information can be reliably transmitted over the channel. For that purpose, we are interested in the mutual information between the input data and output sequence. In the case of a perfect channel, i.e., $I(X;Y) = 1$, we would not have to worry about reliable transmission, and would not need to design a channel encoder and decoder since no information loss occurs. In the case of a completely noisy channel, i.e., $I(X;Y) = 0$, we again would not have to worry since all the information would be lost during transmission and any design would be to no end. In fact, as we will explain in more detail, channel polarization refers to these two extreme situations.

If we assume we make n independent channel uses to transmit information, we know that

$$I(U_1^n; Y_1^n) = I(X_1^n; Y_1^n) = \sum_{i=1}^{n} I(X_i; Y_i) = nI(W) \tag{1.1}$$

where the symmetric capacity of the channel $I(W)$ is defined as

$$I(W) = \sum_{x,y} \frac{1}{2} P(y \mid x) \log \frac{P(y \mid x)}{\frac{1}{2}P(y \mid 0) + \frac{1}{2}P(y \mid 1)}$$

If we apply the chain rule, we obtain

$$I(U_1^n; Y_1^n) = \sum_{i=1}^{n} I(U_i; Y_1^n \mid U_1^{i-1}) = \sum_{i=1}^{n} I(U_i; Y_1^n U_1^{i-1}) \tag{1.2}$$

Therefore instead of the n independent channel uses of W, we can consider from the chain rule perspective n successive uses of the channels $W^{(1,n)} \dots W^{(n,n)}$ given by the transition probabilities $P_1(y_1^n \mid u_1), \dots, P_n(y_1^n u_1^{n-1} \mid u_n)$, respectively.

Following this observation, channel polarization is defined as the case when each of the channels $W^{(i,n)}$ for $i = 1, \dots, n$ converges either to a perfect channel or completely noisy channel. Under this event, we expect the empirical distributions of the mutual information terms in summation (1.2) to satisfy

$$\begin{aligned} &\frac{1}{n} \ \#\{i : I(U_i; Y_1^n U_1^{i-1}) \in [1-\gamma, 1)\} \xrightarrow[n\to\infty]{} I(W) \\ &\frac{1}{n} \ \#\{i : I(U_i; Y_1^n U_1^{i-1}) \in (0, \gamma]\} \xrightarrow[n\to\infty]{} 1 - I(W) \end{aligned} \tag{1.3}$$

for any $\gamma \in (0, 1)$.

In [1], Arıkan discovers a recursive process under which the communication channels $W^{(i,n)}$ exhibit polarization. The idea is exploited to propose a new coding technique, called polar codes, which achieves the symmetric capacity of any B-DMC using a low complexity encoder and decoder. We briefly summarize the process and the structure of the proposed encoder and decoder for polar codes.

To carry the discussion forward, we first need to introduce another channel parameter in addition to the symmetric capacity $I(W)$. The Bhattacharyya parameter of a channel, denoted as $Z(W)$, is defined as

$$Z(W) = \sum_{y \in \mathcal{Y}} \sqrt{P(y \mid 0) P(y \mid 1)}$$

This parameter gives an upper bound to maximum likelihood decoding error probability of a single use of the channel W. Therefore when treated separately, while $I(W)$ stands as a measure of communication rate of the channel, $Z(W)$ stands as a measure of reliability. On the other hand, as we will see shortly, these two parameters are used jointly to prove that channel polarization occurs.

Consider the transformation matrix given as

$$G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{1.4}$$

The corresponding channel configuration is drawn in Figure 1.2 by combining two independent copies of W.

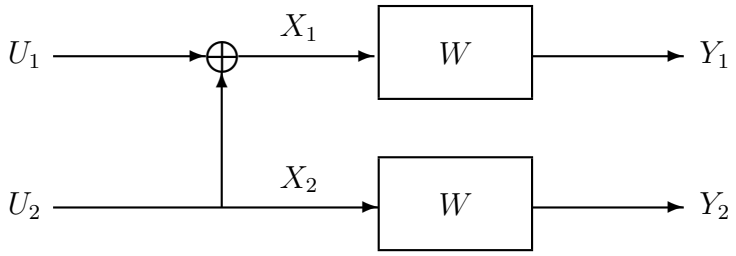

Figure 1.2: Basic channel transformations.

The two successive channels $W^{(1,2)}$ and $W^{(2,2)}$ are characterized by the transformations $W^- : \mathcal{X} \to \mathcal{Y}^2$ and $W^+ : \mathcal{X} \to \mathcal{Y}^2 \times \mathcal{X}$, respectively. Referred as the basic channel transformations, W^- and W^+ can be defined by the following transition probabilities

$$P_{W^-}(y_1 y_2 \mid u_1) = \sum_{u_2 \in \mathcal{X}} \frac{1}{2} P(y_1 \mid u_1 \oplus u_2) P(y_2 \mid u_2) \tag{1.5}$$

$$P_{W^+}(y_1 y_2 u_1 \mid u_2) = \frac{1}{2} P(y_1 \mid u_1 \oplus u_2) P(y_2 \mid u_2) \tag{1.6}$$

The following properties related to the above transformations are derived in [1]:

1. The mutual information is preserved:

$$I(W^-) + I(W^+) = 2I(W)$$

2. The overall reliability is improved:

$$Z(W^-) + Z(W^+) \leq Z(W)$$

3. While the channel W^+ is improved, the channel W^- is worsened:

$$\begin{aligned} I(W^-) &\leq I(W) \leq I(W^+) \\ Z(W^-) &\geq Z(W) \geq Z(W^+) \end{aligned}$$

4. The evolution of Z(W) satisfies

$$\begin{aligned} Z(W^-) &\leq 2Z(W) - Z(W)^2 \\ Z(W^+) &= Z(W)^2 \end{aligned}$$

where equality holds in the first line when W is a binary erasure channel.

5. The following bounds can be derived:

$$\begin{aligned} I(W) \ + \ Z(W) \ &\geq 1 \\ I(W)^2 + \ Z(W)^2 &\leq 1 \end{aligned}$$

where equality holds in the first inequality when W is a binary erasure channel.

6. Given a binary erasure channel W_{BEC}, the channels W^-_{BEC} and W^+_{BEC} are also binary erasure channels.

The third property confirms that we are in the right way to polarization. The idea now is to apply the same basic channel transformations to the channels W^- and W^+. As a result, four channels W^{--}, W^{-+}, W^{+-}, and W^{++} are obtained. However in general, we are no longer able to compare the parameters of these four channels in terms of rate and reliability, except the knowledge that the channel W^{++} is the best one and the channel W^{--} is the worst one. Nevertheless, recall that we require in equations (1.3) convergence to polarized channels with the sequence length. Hence, we keep applying the transformations to the obtained $\pm$ channels. To analyze the convergence properties of this recursion, the polarization process is defined.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. The random sequence $B_1, \ldots, B_\ell$ is drawn i.i.d according to a Bernoulli distribution with probabilities equal to $\frac{1}{2}$. Let $\mathcal{F}_\ell$ be the σ-algebra generated by this Bernoulli sequence. Given a channel W, the random sequence of channels $\{W_\ell\}$ is defined for $\ell \geq 0$ as

$$W_\ell = \begin{cases} W & \text{if} \quad \ell \ = 0 \\ W^-_{\ell-1} & \text{if} \quad B_\ell = 0 \\ W^+_{\ell-1} & \text{if} \quad B_\ell = 1 \end{cases}$$

In the sequel, the random processes $I_\ell = I(W_\ell)$ and $Z_\ell = Z(W_\ell)$ are defined and the next properties are proved in [1]:

7. By definition, the properties 1-6 hold for I_ℓ and Z_ℓ at each ℓ.

8. The process $\{I_\ell, \mathcal{F}_\ell\}$ is a bounded martingale on the interval $[0, 1]$, i.e.,

$$\begin{aligned} \mathbb{E}[I_\ell \mid I_1, \ldots, I_{\ell-1}] &= I(W^-_{\ell-1}) \ \mathbb{P}\big[I_\ell = I(W^-_{\ell-1})\big] + I(W^+_{\ell-1}) \ \mathbb{P}\big[I_\ell = I(W^+_{\ell-1})\big] \\ &= I_\ell \end{aligned}$$

9. The process $\{Z_\ell, \mathcal{F}_\ell\}$ is a bounded supermartingale on the interval $[0, 1]$, i.e.,

$$\begin{aligned} \mathbb{E}[Z_\ell \mid Z_1, \ldots, Z_{\ell-1}] &= Z(W^-_{\ell-1}) \ \mathbb{P}\big[Z_\ell = Z(W^-_{\ell-1})\big] + Z(W^+_{\ell-1}) \ \mathbb{P}\big[Z_\ell = Z(W^+_{\ell-1})\big] \\ &\leq Z_\ell \end{aligned}$$

10. The process $\{Z_\ell, \mathcal{F}_\ell\}$ converges a.s. to a $\{0, 1\}$ valued random variable Z_∞ since

$$\begin{aligned} & \mathbb{E}[|Z_{\ell+1} - Z_\ell|] \xrightarrow[\ell\to\infty]{} 0 \quad \& \quad \mathbb{P}\big[Z_{\ell+1} = Z_\ell^2\big] = \frac{1}{2} \\ \Rightarrow \quad & \mathbb{E}[|Z_{\ell+1} - Z_\ell|] \geq \frac{1}{2} \ \mathbb{E}[|Z_\ell \, (1 - Z_\ell)|] \xrightarrow[\ell\to\infty]{} 0 \end{aligned}$$

11. The process $\{I_\ell, \mathcal{F}_\ell\}$ converges a.s. to a random variable I_∞ such that

$$\mathbb{E}[|I_\infty|] = I_0$$

where I_∞ takes values a.s in $\{0, 1\}$ since

$$\begin{aligned} & I_\ell + Z_\ell \geq 1 \quad \& \quad I_\ell^2 + Z_\ell^2 \leq 1 \quad \& \quad Z_\infty \in \{0, 1\} \\ \Rightarrow \quad & I_\infty + Z_\infty = 1 \end{aligned}$$

The last property proves that the recursive application of the basic channel transformations lead to channel polarization. The next theorem states this result.

Theorem 1.1. [1] *For any B-DMC W, the sequence of channels $\{W_\ell\}$ polarizes such that*

$$\begin{aligned} \frac{1}{2^\ell} \#\{i : I(W^{(i,2^\ell)}) \in [1-\gamma, 1)\} &\xrightarrow[\ell\to\infty]{} I(W) \\ \frac{1}{2^\ell} \#\{i : I(W^{(i,2^\ell)}) \in (0, \gamma]\} &\xrightarrow[\ell\to\infty]{} 1 - I(W) \end{aligned}$$

for fixed $\gamma \in (0, 1)$ and $i = 1, \ldots, 2^\ell$.

Using the idea of channel polarization, polar codes are proposed as a new coding technique. Namely, a polar code (N, R) of block length N and rate R communicates information only on the $\lfloor NR \rfloor$ good channels whose mutual information values are close to 1 among all the synthesized $W^{(i,N)}$ channels with $i = 1, \ldots, N$, and simply transmits randomly fixed bits known to both the encoder and decoder over the other bad channels. We explain the details on how polar codes can be constructed in the next section. Below, we first summarize the structure of the proposed channel encoder and decoder and the expected code performance [1].

Encoder:

12. The recursion can be applied through the channel transformation matrix

$$G_N = G^{\otimes N}$$

where $\otimes$ denotes the Kronecker product and G is defined in (1.4). Given the data sequence u_1^N, the input sequence x_1^N is computed from

$$x_1^N = u_1^N B_N G_N$$

where B_N is a permutation matrix known as bit-reversal.

13. Depending on the direction of polarization of the channels, the data sequence $u_1 \ldots u_N$ indexes are split into two sets before transmission. The first one includes the indexes of the data to be transmitted on the good channels, and is referred as the information set $\mathcal{A}_N$. The remaining one is the set $\mathcal{A}_N^c$ of the indexes corresponding to the frozen bits to be transmitted on the bad channels.

14. For symmetric channels the frozen bits can be selected in a deterministic way as they don't affect the code performance.

15. The above encoder can be implemented in $\Theta(N \log N)$ complexity.

Decoder:

16. We want to decode the output of the N channels defined by the transition probabilities $P_1(y_1^N \mid u_1), \ldots, P_N(y_1^N u_1^{n-1} \mid u_N)$. Therefore, we can see that to estimate the channel input $\hat{u}_i$ in this chain we need the correct estimates $\hat{u_1}, \ldots, \hat{u}_{i-1}$ of the previous channel inputs. This form suggests the use of a successive cancellation decoder.

17. Arıkan proposes a similar decision rule to maximum likelihood decoder for the non-frozen bits

$$\mathcal{D}(y_1^N, u_1^{i-1}) = \begin{cases} 0 & \text{if } \dfrac{P_i(y_1^N u_1^{i-1} \mid 0)}{P_i(y_1^N u_1^{i-1} \mid 1)} \geq 1 \\ 1 & \text{otherwise} \end{cases}$$

For the frozen bits, the decoder can directly set them since their values are already available.

18. The successive cancellation decoder using the above decision rule can be implemented in $\Theta(N \log N)$ complexity.

Code Performance:

19. Let $P_e(N, R)$ denotes the best achievable block error probability under successive cancellation decoding over the ensemble of all possible choices of the set $\mathcal{A}_N^c$. Then,

$$P_e(N, R) \leq \sum_{i \in \mathcal{A}_N} Z(W^{(i,N)}) \tag{1.7}$$

1.2 Project Motivation

Although channel polarization does occur and polar codes achieving the symmetric capacity of any B-DMC can be implemented using low complexity encoders and decoders, the rate of polarization is crucial to make polar codes part of any real application. We need to ensure that channel polarization takes place fast enough as, in practice, the performance of polar codes with finite block lengths are important. We first state the best result known on the rate of channel polarization.

Theorem 1.2. [4] *Given any rate $R \geq 0$ such that $R \leq I(W)$ and a constant $\beta \leq \frac{1}{2}$, consider the polar code (N, R) of block length N. Then,*

$$P_e(N, R) = \Theta(2^{-N^\beta}) \tag{1.8}$$

A closely tied problem to the rate of channel polarization is the polar code construction problem. In property 13, we defined the information set $\mathcal{A}_N$. In fact, polar code construction is nothing but the choice of the indexes in this set. For a fixed block length N, Arıkan suggests to construct polar codes such that the error probability expression in (1.7) is minimized for a given threshold $\eta \in (0, 1)$. More precisely,

$$\mathcal{A}_N(W, \eta) = \{i : Z(W^{(i,N)}) \leq \eta\} \tag{1.9}$$

We know how to compute the $Z(W^{(i,N)})$ efficiently only when we have a binary erasure channel. However, by the equality condition mentioned in property 4, we see that the binary erasure channel gives an upper bound to the Bhattacharyya parameters obtained after applying one level of the recursion for a fixed value of the Bhattacharyya parameter. Moreover, from to the statement of property 6, we know that a binary erasure channel remains so after any level of the recursion. Hence, we deduce that the binary erasure channel W_{BEC} represents an extremal channel among all B-DMC's. Given $Z(W) = Z(W_{\mathsf{BEC}})$, the parameters $Z(W^{(i,N)})$ can be upper bounded as

$$Z(W^{(i,N)}) \leq Z(W_{\mathsf{BEC}}^{(i,N)}) \quad \forall \ i = 1, \ldots, N$$

and consequently,

$$\mathcal{A}_N(W_{\mathsf{BEC}}, \eta) \subseteq \mathcal{A}_N(W, \eta)$$

Similarly, one can show that the binary symmetric channel W_{BSC} gives a lower bound to the Bhattacharyya parameters of the channels after applying one level of the recursion. When $Z(W) = Z(W_{\mathsf{BSC}})$, we have

$$Z(W_{\mathsf{BSC}}^{(i,2)}) \leq Z(W^{(i,2)}) \quad \text{for} \ \ i = 1, 2$$

Given a binary symmetric channel W_{BSC}, while the channel W_{BSC}^{-} is a binary symmetric channel, the channel W_{BSC}^{+} is not. Hence at this point, we don't reach an extremality result as strong as the binary erasure channel's one.

To sum up, these attribute special importance to the binary erasure channel and the binary symmetric channel in the design of polar codes. For instance in [5], Arıkan carries a performance comparison with Reed-Muller codes by using polar codes over binary erasure channels.

On the other hand, we know that $I(W)$ and $Z(W)$ can be used interchangeably to state arguments about polar codes. By intuition, we suspect other channel parameters might be helpful to better understand polar codes. For that reason, the function $E_0(\rho, W)$ defined by Gallager [2] as a part of the random coding exponent draw our attention. Both the random coding exponent and $E_0(\rho, W)$ can be viewed as more general channel parameters. Actually, $I(W)$ and $Z(W)$ can be derived as special cases of $E_0(\rho, W)$:

- $I(W)$ is the slope of the function $E_0(\rho, W)$ evaluated at $\rho = 0$.
- $E_0(\rho, W)|_{\rho=1} = \log \dfrac{2}{1 + Z(W)}$

Following all these observations, the main motivation behind this project is to study the behavior of $E_0(\rho, W)$ and the random coding exponent from the aspect of polarization and polar codes. For that purpose, we analyze their properties and we inquire how $E_0(\rho, W)$ is affected by the basic channel transformations. In Theorem 4.5 and Theorem 4.6, we show that the binary erasure channel and the binary symmetric channel are also extremal in the evolution of $E_0(\rho, W)$ under the basic channel transformations.

To further motivate this work with an application, we provide a discussion related to the compound capacity of polar codes. A recent work [6] on this subject formalize upper and lower bounds on this capacity. The lower bound is based on the same idea that among all polarization processes Z_n with a given initial value of the Bhattacharyya parameter, the process corresponding to the binary erasure channel is extremal and constitutes an upper bound to all other processes. This application is introduced as an evidence on the importance of the extremal property related to the Bhattacharyya parameter. As in this application, we believe the extremality relations we derive for $E_0(\rho, W)$ might be used to arrive at new results on polar codes.

1.3 Outline

In Chapter 2, we give a formal definition of the random coding exponent and the function $E_0(\rho, W)$. Then, we summarize their properties and mention an alternative description of the function $E_0(\rho, W)$ in terms of Rényi's entropy functions. We also provide two applications in information theory.

In Chapter 3, we discuss the compound capacity of symmetric channels. We provide a similar proof to the Shannon's proof of the random coding theorem, and based on this proof, we show that linear codes achieving this compound capacity exist. Subsequently, we explain existing upper and lower bounds on the compound capacity of polar codes.

In Chapter 4, we present the main results. We explore the evolution of $E_0(\rho, W)$ under the basic channel transformations in detail. We show that for a given $E_0(\rho, W)$, the binary erasure channel and the binary symmetric channel are extremal with respect to $E_0(\rho, W^-)$, and $E_0(\rho, W^+)$. Then, we describe $E_0(\rho, W^-)$ and $E_0(\rho, W^+)$ in terms of Rényi's entropy functions. Based on this representation, we conjecture an inequality between $E_0(\rho, W)$, $E_0(\rho, W^-)$, and $E_0(\rho, W^+)$. Lastly in Section 4.4, independent of channel polarization, we search for more cases where the binary erasure channel and the binary symmetric channel are extremal.

The final chapter is the Conclusions. The rest of the report contains the Appendices.

Chapter 2

Random Coding Exponent and E_0

In this chapter, we discuss the random coding exponent, denoted as $E_r(R)$, and the function E_0 which is essential in its characterization. We start with the mathematical definition. Then, we summarize the properties and identify the connections with channel polarization and polar codes. We introduce a result by Arıkan and Telatar [3], which reveals that the binary erasure channel and the binary symmetric channel are $E_r(R)$ extremal. We also provide an alternative description of E_0 that uses Rényi's entropy functions. These constitute an important basis for the work we present in Chapter 4. Finally, we conclude this chapter by two applications where $E_r(R)$ and E_0 appear. The first one, due to Gallager [2], is in the upper bound to block error probability used to prove the well-known channel coding theorem. The second one, due to Arıkan [7], is in the lower bound to computational complexity of sequential decoding.

2.1 Definition and Properties

Definition 2.1. [2] Let a discrete memoryless channel W with input random variable $X \sim Q(x)$ and output random variable Y have input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probabilities $P(y \mid x)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Consider the function $E_r(R, Q)$ defined as

$$E_r(R, Q) = \max_{\rho \in [0,1]} \{E_0(\rho, Q) - \rho R\}$$

where

$$E_0(\rho, Q) = -\log \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} Q(x) P(y \mid x)^{\frac{1}{1+\rho}} \right]^{1+\rho} \tag{2.1}$$

Then, the random coding exponent is defined as

$$E_r(R) = \max_{Q} E_r(R, Q) \tag{2.2}$$

From the above construction, we see that the function $E_0(\rho, Q)$ is central in the behavior of the random coding exponent. Before we introduce the properties, we first restrict the analysis to inputs with binary uniform distribution to be consistent with

the basic channel transformations we defined in (1.5) and (1.6). This is reflected in the above notations by replacing $Q \to W$ to emphasize the dependence on the channel. Therefore we are no longer interested in the maximization in (2.2), and the equation in (2.1) becomes

$$E_0(\rho, W) = -\log \sum_{y \in \mathcal{Y}} \left[\frac{1}{2} P(y \mid 0)^{\frac{1}{1+\rho}} + \frac{1}{2} P(y \mid 1)^{\frac{1}{1+\rho}} \right]^{1+\rho} \tag{2.3}$$

Figure 2.1 shows the $E_0(\rho, W)$ versus ρ curves of a binary erasure channel (BEC) and a binary symmetric channel (BSC).

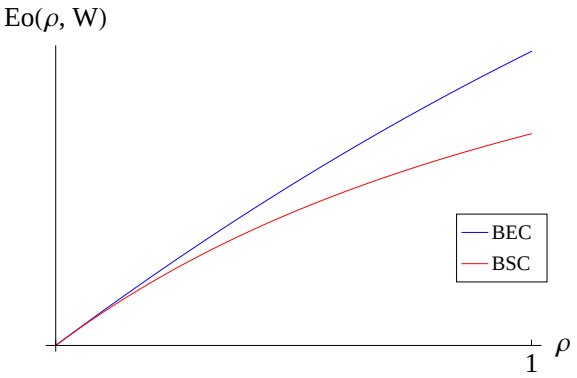

Figure 2.1: $E_0(\rho, W)$ vs ρ plot for BEC and BSC with $I(W) = 0.5$.

Theorem 5.6.3 in [2] summarizes the properties of $E_0(\rho, W)$ with respect to the variable ρ. For $\rho \geq 0$, $E_0(\rho, W)$ is a positive, concave increasing function in ρ. Moreover, one can easily derive the following relationships which show how the parameters $I(W)$ and $Z(W)$ are related to $E_0(\rho, W)$:

- $E_0(\rho, W)\Big|_{\rho=0} = 0$
- $E_0(\rho, W)\Big|_{\rho=1} = \log \frac{2}{1+Z(W)} \quad \Rightarrow \quad Z(W) = 2 \times 2^{-\frac{E_0(\rho,W)}{\rho}\Big|_{\rho=1}} - 1$
- $\frac{\partial}{\partial \rho} E_0(\rho, W)\Big|_{\rho=0} = I(W) \quad \Rightarrow \quad \frac{E_0(\rho, W)}{\rho}\Big|_{\rho=0} = I(W)$

By the concavity of the function $E_0(\rho, W)$, for R in the range

$$\frac{\partial E_0(\rho, W)}{\partial \rho}\Big|_{\rho=1} \leq R \leq \frac{\partial E_0(\rho, W)}{\partial \rho}\Big|_{\rho=0}$$

the maximization of $E_r(R, W)$ over $\rho \in [0, 1]$ can be described in terms of the following parametric equations

$$\begin{aligned} R(\rho, W) &= \frac{\partial}{\partial \rho} E_0(\rho, W) \\ E_r(\rho, W) &= E_0(\rho, W) - \rho \frac{\partial}{\partial \rho} E_0(\rho, W) \end{aligned} \tag{2.4}$$

Theorem 5.6.4 in [2] formalizes the properties of $E_r(R, W)$ and provides a graphical interpretation of the $E_r(R, W)$ versus $R(\rho, W)$ curve. For each value of $\rho \in [0, 1]$, $R = R(\rho, W)$ is a constant and the $E_r(R, W)$ versus R plot is a line with slope $-\rho$ which intersects the $E_r(R, W)$ axis at $E_0(\rho, W)$. Therefore, the $E_r(R, W)$ curve can be generated as the lowest upper bound to all the lines plotted at each $\rho \in [0, 1]$. As a result, $E_r(R, W)$ is a positive, convex decreasing function in R for any given B-DMC W.

On the other hand, another interesting property of both $E_0(\rho, W)$ and $E_r(\rho, W)$ with respect to the channel W is pointed in [3]. The authors show that given a channel W, a binary erasure channel W_{BEC}, and a binary symmetric channel W_{BSC} with the same rate, i.e., such that the equality

$$R(\rho, W) = R(\rho, W_{\mathsf{BEC}}) = R(\rho, W_{\mathsf{BSC}})$$

holds for a fixed value of $\rho \in [0, 1]$, then

$$E_0(\rho, W_{\mathsf{BEC}}) \leq E_0(\rho, W) \leq E_0(\rho, W_{\mathsf{BSC}})$$
$$E_r(\rho, W_{\mathsf{BEC}}) \leq E_r(\rho, W) \leq E_r(\rho, W_{\mathsf{BSC}})$$

The proof follows from the fact that, for a fixed ρ value, one can write $R(\rho, W)$ and $\exp\{E_0(\rho, W)\}$ as the expected value of functions of a random variable taking values in $[0, 1]$. In fact, the binary erasure channel whose variable takes only the extremal values $\{0, 1\}$ and the binary symmetric channel whose variable takes a constant value are special cases of this random variable. Then, a convexity analysis combined with the Jensen's inequalities lead to the final result.

Figures 2.2 and 2.3 show the $E_0(\rho, W)$ versus $R(\rho, W)$ and $E_r(R, W)$ versus $R(\rho, W)$ curves, respectively, of a binary erasure channel and a binary symmetric channel parametrized in terms of their channel parameters for a fixed value of ρ.

2.2 E_0 Description by Rényi's Entropy Functions

The previous section revealed that the symmetric capacity $I(W)$ and the Bhattacharyya parameter $Z(W)$ can be derived from the function $E_0(\rho, W)/\rho$ evaluated on the limit $\rho \to 0$ and at $\rho = 1$, respectively. In this section, we provide an alternative description of $E_0(\rho, W)/\rho$ using the concept of Rényi's entropy functions. The importance lies in the fact that this gives an interpretation to $E_0(\rho, W)/\rho$ as a general measure of information.

Rényi's entropy function of order α of a discrete random variable $X \sim P(x)$ is defined in [8] as

$$H_\alpha(X) = \frac{\alpha}{1-\alpha} \log \left(\sum_x P(x)^\alpha \right)^{\frac{1}{\alpha}}$$

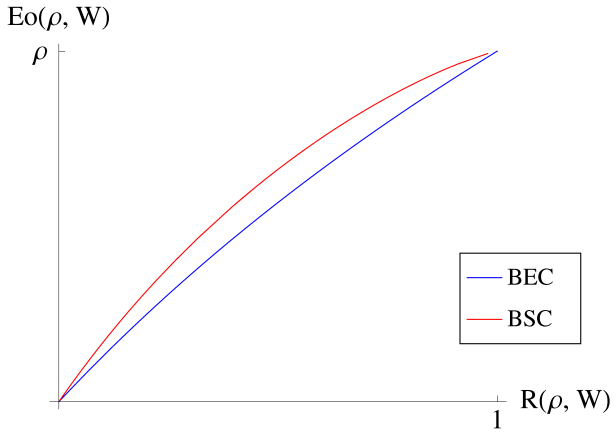

Figure 2.2: Parametric $E_0(\rho, W)$ vs $R(\rho, W)$ curve for BEC and BSC with fixed ρ

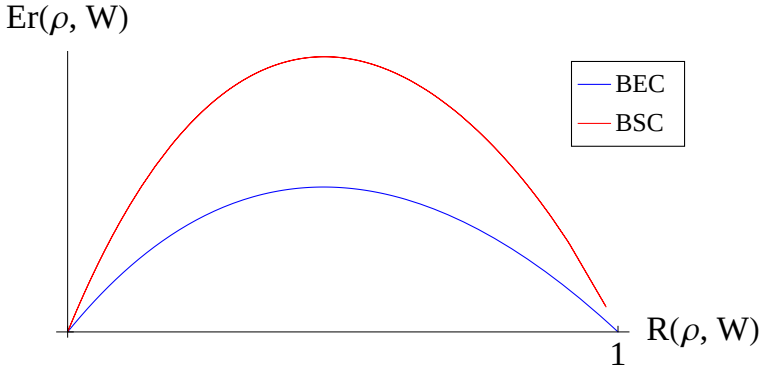

Figure 2.3: Parametric $E_r(\rho, W)$ vs $R(\rho, W)$ curve for BEC and BSC with fixed ρ

Now, we seek for a possible extension of the above definition to define a quantity similar to the conditional entropy function. In this effort, we note that different definitions exist in the literature. As one suitable for the study of channel polarization, we use the Rényi's conditional entropy function of order α of a discrete random variable X given Y with joint distribution $P(x, y)$ defined in [9] as

$$\begin{aligned} H_\alpha(X \mid Y) &= \frac{\alpha}{1-\alpha} \log \sum_y \left(\sum_x P(x,y)^\alpha \right)^{\frac{1}{\alpha}} \\ &= H_\alpha(X) + \frac{\alpha}{1-\alpha} \log \sum_y \left(\sum_x Q(x) P(y \mid x)^\alpha \right)^{\frac{1}{\alpha}} \end{aligned}$$

where $Q(x) = \dfrac{P(x)^\alpha}{\sum_x P(x)^\alpha}$ is a probability distribution.

If we assume uniform input distribution and let $\alpha = \frac{1}{1+\rho}$, we get

$$H_{\frac{1}{1+\rho}}(X) = \frac{1}{\rho} \log \left(\sum_x P(x)^{\frac{1}{1+\rho}} \right)^{\frac{1}{1+\rho}} \tag{2.5}$$

$$H_{\frac{1}{1+\rho}}(X \mid Y) = H_{\frac{1}{1+\rho}}(X) + \frac{1}{\rho} \log \sum_y \left(\sum_x P(x)P(y \mid x)^{\frac{1}{1+\rho}} \right)^{1+\rho} \tag{2.6}$$

From the definition of $E_0(\rho, W)$ in (2.3), we deduce

$$\frac{E_0(\rho, W)}{\rho} = H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(X \mid Y) \tag{2.7}$$

The quantity in the RHS of (2.7) is called as the mutual information of order $\frac{1}{1+\rho}$ in [9]. Moreover, the following properties are proved

- $\lim_{\alpha \to 1} H_\alpha(X) = H(X \mid Y)$
- $H_\alpha(X \mid Y) \leq H_\alpha(X)$, i.e "conditioning reduces entropy" is valid for Rényi's entropy function, as it is in the Shannon entropy case.
- $\frac{E_0(\rho, W)}{\rho}$ is a decreasing function in ρ.

2.3 Applications

We close this chapter with two applications where $E_r(R)$ and E_0 play a critical role. We should mention that the purpose of this section is only to give the main idea about how these were used in information theoretic problems. Hence, we avoid digressing into different topics.

2.3.1 Upper Bound to Block Error Probability

Reliable transmission of information from one point to another point is the ultimate goal of communication systems. To achieve this objective, channel encoders use appropriate coding schemes to construct redundant messages, such that upon reception, the corresponding channel decoders can overcome the effects of noise introduced during transmission. In the analysis of this problem, different criteria and trade-offs in-between are considered. The rate of transmission, the decoding error probability, the encoders and decoders complexities are among the most important ones.

To prove fundamental results, channel block codes, simpler to treat, are used as a mean. The block decoding error probability, denoted as P_e, is a criteria used to

assess performance of block codes. Furthermore, in the analysis of P_e, random ensembles of block codes received particular attention. If the average block decoding error probability, denoted as $P_{e,avg}$, is expected to satisfy desirable properties over the ensemble, then we can find a specific realization that will do at least as good as the average. In chapter 5 of [2], the attempts to upper bound $P_{e,avg}$ resulted in the Channel Coding Theorem (5.6.2), and the definition of the random coding exponent. Broadly speaking, this theorem states that for a DMC, and a fixed rate $R > 0$, the average block decoding error probability of an ensemble of block codes with codewords of length N, can be upper bounded as

$$P_{e,avg} \leq \exp\{-NE_r(R)\} \tag{2.8}$$

In fact, for any $\rho \in [0, 1]$, we have the following more general bound

$$P_{e,avg} \leq \exp\{-N\left(E_0(\rho, W) - \rho R\right)\} \tag{2.9}$$

The bound in (2.8) follows by taking the ρ value that gives the tightest bound in (2.9).

Therefore, $E_r(R)$ establishes the compromise between $P_{e,avg}$, the block length N, and the communication rate R. Obviously, if $E_r(R) > 0$, then $P_{e,avg}$ vanishes exponentially with increasing block lengths. We have already mentioned the properties of $E_r(R)$. These properties were used to prove the Noisy Channel Coding Theorem [2] fundamental in the interpretation of the existing trade-off. The theorem states that $E_r(R) > 0$ for fixed rates below channel capacity, i.e. $R \leq C$. Hence by increasing the code block length, we can make the block decoding error probability arbitrarily small.

We last want to mention, without going into details, an important step in the above derivations. The error probability can be computed by taking the union of the probabilities of all the events which cause the decoder to make an error. Let us define these error events as E_i with $i = 1, \ldots, M$. Then, as a consequence of the union bound,

$$P\left(\bigcup_i E_i\right) \leq \left[\sum_i P(E_i)\right]^{\rho}$$

holds for any $\rho \in [0, 1]$. Therefore, the variable ρ is introduced to improve the bound when the summation of the probabilities are larger than 1 at the cost of providing less tight bounds otherwise.

2.3.2 Lower Bound to Computational Complexity of Sequential Decoding

Assume an input sequence X is transmitted through a channel W, and the output sequence Y is received. A sequential decoder can be described as a device which, based on the received value, keeps guessing which particular input was transmitted until the correct decision is made. The number of guesses made during this

procedure, and the number of computations performed by the decoder are parallel quantities. They both depend on the order in which guesses are made. This order, in turn, is determined by a function $G(x \mid y)$ called a guessing function. Therefore, the computational complexity of sequential decoding can be expressed in terms of the random variable $G(X \mid Y)$.

In [10], Massey considers the guessing problem standalone, and defines the form of an optimal guessing function. According to the paper, the average number of guesses is minimized by guessing the value of the random variable X given Y in decreasing order of conditional probabilities. In [7], Arıkan considers sequential decoders which have arbitrary guessing functions. The following lower bound to the moments of computational complexity of sequential decoding is given

$$\mathbb{E}[G(X \mid Y)^{\rho}] \geq (1 + NR)^{-\rho} \exp\{N\left(\rho R - E_0(\rho, W)\right)\}$$

for $\rho > 0$. Note the difference between the above exponent and $E_r(R)$, which involved a maximization over $\rho \in [0, 1]$ in the parametric equations (2.4).

As in the previous application, a trade-off between the rate, the block length and this time the computational complexity is discovered. The critical value $E_0(\rho, W)/\rho$, called the cut-off rate, imposes a limit on the rate R. Above this limit, as the block length N is increased, infinitely many computations need to be performed by the sequential decoder. Hence, complexity is unbounded.

Finally, let us consider the two applications within the same framework. On one side, the random channel coding theorem tells that one can communicate at rates up to channel capacity with arbitrarily small error probabilities. On the other hand, the particular choice of the decoder as a sequential decoder restricts the communication rate to the cut-off rate of the channel, since we know that $E_0(\rho, W)/\rho \leq I(W)$ for $\rho > 0$ from the previous section.

Chapter 3

Compound Capacity of Symmetric Channels

In the previous chapters, all the arguments were based on the assumption that we are given a specific channel over which we want to communicate. However in some cases, only a partial knowledge on the communication channel is available. For instance, we might just know the possible range of values the mutual information between the input and output of the channel takes. In other cases, we might want to design codes that perform well not only in a particular channel but also in other channels . Both of these situations justify the need to analyze performance of codes constrained to broader type of channels. We first introduce the notion of compound capacity which extends the definition of capacity for a single channel to a class of channels.

The compound capacity of a class of channels is given by [11]

$$C(\mathcal{W}) = \max_{Q(x)} \inf_{W \in \mathcal{W}} I(Q, W)$$

where $\mathcal{W}$ represents the class of channels, $Q(x)$ is the input distribution, and $I(Q, W)$ is the corresponding mutual information between the input and output of the channel.

In general, $C(\mathcal{W})$ is smaller than the infimum of any $I(Q, W)$ in $\mathcal{W}$. However, we can restrict the analysis to binary memoryless symmetric (BMS) channels and define a symmetric compound capacity by the formula

$$\mathcal{I}(\mathcal{W}) = \min_{W \in \mathcal{W}} I(W) \tag{3.1}$$

since the uniform input distribution corresponds to the maximizing input distribution. Conforming to our previous notation, $I(W)$ equals the symmetric capacity of the channel.

In the rest of this chapter, we start by proving that linear codes that achieve the symmetric compound capacity given in (3.1) exist. Then, we discuss existing results for polar codes. We should mention that linear codes are not chosen at random. Both

linear codes and polar codes can be constructed with an encoder that transforms the input u_1^n to a codeword $x_1^n = u_1^n G_n$ with a generator matrix G_n. The difference is on the structure of the generator matrix, i.e., on how the elements of G_n are chosen. For linear codes, the elements are selected arbitrarily from a finite field, such as GF(2) when the input is binary. Recall that for polar codes, the rows are selected in concordance with channel polarization phenomenon.

3.1 Linear Codes

Definition 3.1. Given a field $(\mathcal{X}, +, .)$, we say that a code $\mathcal{C} \subset \mathcal{X}^n$ is linear if it is a vector space. Therefore,

$$\forall a, b \in \mathcal{X}, \;\; \forall \bar{x}, \bar{y} \in \mathcal{C} \quad \Rightarrow \quad a\bar{x} + b\bar{y} \in \mathcal{C}$$

The channel coding theorem for linear codes states that capacity achieving linear codes exist. The proof simply relies on the Shannon's proof of the random coding theorem. Here, we follow a similar approach. We first note the $\mathcal{I}(\mathcal{W})$ can be achieved by a random coding argument similar to the single channel case.

Theorem 3.2. *Let $\mathcal{W}$ be a set of symmetric channels with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, uniform input distribution and transition probabilities $P_{\mathsf{w}}(y \mid x)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ for each $W \in \mathcal{W}$. Then, there exists a block code of block length n and rate $R \geq 0$ such that $R \leq \min_{W \in \mathcal{W}} I(X;Y)$ and for any $\epsilon > 0$ the average block decoding error probability $P_{e,avg} < \epsilon$ with $\epsilon \to 0$ as $n \to \infty$.*

Before we start proving the theorem, we introduce the concept of strong typicality we use in the proof.

Definition 3.3. The strongly typical set $\mathsf{T}^{n,\epsilon}_{P_{\mathsf{z}}(z)}$ of sequences $z_1 \dots z_n \in \mathcal{Z}^n$ with respect to the distribution $P_{\mathsf{z}}(z)$ can be defined as

$$\mathsf{T}^{n,\epsilon}_{P_{\mathsf{z}}(z)} = \left\{ \begin{array}{l} (z_1 \dots z_n) \in \mathcal{Z}^n : \\ \forall \zeta \in \mathcal{Z}, \quad \frac{1}{n} \# \{i : z_i = \zeta\} = \left\{ \begin{array}{ll} 0 & \text{if } P_{\mathsf{z}}(\zeta) = 0 \\ P_{\mathsf{z}}(\zeta) \pm \epsilon & \text{otherwise} \end{array} \right. \end{array} \right\}$$

Lemma 3.4. *Let $Z_1 \dots Z_n$ be a sequence of random variables drawn identically independently according to the distribution $P_{\mathbf{z}}(z_1, \dots, z_n) = \prod_i P_{\mathsf{z}}(z_i)$. Given the distribution $Q_{\mathbf{z}}(z_1, \dots, z_n)$, we have*

$$\mathbb{P}\left((z_1 \dots z_n) \in \mathsf{T}^{n,\epsilon}_{Q_{\mathsf{z}}(z)} \right) = 2^{-n(\mathsf{D}(Q||P) \pm \epsilon)}$$

where $\mathsf{D}(Q \;||\; P) = \sum_z Q_{\mathsf{z}}(z) \log \dfrac{Q_{\mathsf{z}}(z)}{P_{\mathsf{z}}(z)}$ is the Kullback-Leibler divergence between $Q_{\mathsf{z}}(z)$ and $P_{\mathsf{z}}(z)$.

Proof of Lemma 3.4. We define $\bar{z} = (z_1 \dots z_n)$. Then

$$\begin{aligned}
\mathbb{P}\left(\bar{z} \in \mathsf{T}^{n,\epsilon}_{Q_{\mathsf{z}}(z)}\right) &= \sum_{\bar{z}\in\mathsf{T}^{n,\epsilon}_{Q}} P_{\mathsf{z}}(z_1)\dots P_{\mathsf{z}}(z_n) \\
&= \sum_{\bar{z}\in\mathsf{T}^{n,\epsilon}_{Q}} \prod_{z\in\mathcal{Z}} P_{\mathsf{z}}(z)^{nQ_{\mathsf{z}}(z)} \\
&= \sum_{\bar{z}\in\mathsf{T}^{n,\epsilon}_{Q}} 2^{n\sum_{z\in\mathcal{Z}} Q_{\mathsf{z}}(z)\log P_{\mathsf{z}}(z)} \\
&\overset{(1)}{=} 2^{n\left(H_Q(Z)\pm\epsilon\right)} \quad 2^{n\sum_{z\in\mathcal{Z}} Q_{\mathsf{z}}(z)\log P_{\mathsf{z}}(z)} \\
&= 2^{-n(D(Q||P)\pm\epsilon)}
\end{aligned}$$

where (1) follows by the strong asymptotic equipartition property. □

Proof of Theorem 3.2. We assume that for each message $m = 1, \dots, M$ the encoder generates the codewords using the function $Enc : \{1, \dots, M\} \to \mathcal{X}^n$

$$Enc(m) = \bar{x}_m = \{x_1 \dots x_n\}_m$$

Given the output sequence $\bar{y} = y_1 \dots y_n$ i.i.d from $p_{\mathsf{w}_o}(y)$, the decoder makes a decision using the function $Dec : \mathcal{Y}^n \to \{1, \dots, M\} \cup 0$

$$Dec(\bar{y}) = \begin{cases} m & \text{if m is the unique message such that} (Enc(m), \bar{y}) \in \bigcup\limits_{W} \mathsf{T}^{n,\epsilon}_{P_{\mathsf{x}}(x)P_{\mathsf{w}}(y|x)} \\ 0 & \text{otherwise} \end{cases}$$

Let the block decoding error probability of a message m be $P_{e,m}$. We note that

$$\mathbb{E}[P_{e,avg}] = \sum_{m} \frac{1}{M}\mathbb{E}[P_{e,m}] = \mathbb{E}[P_{e,m}]$$

since by symmetry $\mathbb{E}[P_{e,1}] = \dots = \mathbb{E}[P_{e,M}]$.

The decoder makes an error if and only if

- $(\bar{x}_m, \bar{y}) \notin \mathsf{T}^{n,\epsilon}_{P_{\mathsf{x}}(x)P_{\mathsf{w}_o}(y|x)}$
- For any $m' \neq m$, $(\bar{x}_{m'}, \bar{y}) \in \bigcup\limits_{W} \mathsf{T}^{n,\epsilon}_{P_{\mathsf{x}}(x)P_{\mathsf{w}}(y|x)}$

Hence,

$$\mathbb{E}[P_{e,m}] \leq \mathbb{P}\left((\bar{x}_m, \bar{y}) \notin \mathsf{T}^{n,\epsilon}_{P_{\mathsf{x}}(x)P_{\mathsf{w}_o}(y|x)}\right) + \sum_{m'\neq m} \mathbb{P}\left((\bar{x}_{m'}, \bar{y}) \in \bigcup_{W} \mathsf{T}^{n,\epsilon}_{P_{\mathsf{x}}(x)P_{\mathsf{w}}(y|x)}\right)$$

Moreover, we can deduce that $P_{\mathbf{x},\mathbf{y}}(\bar{x}_{m'}, \bar{y}) = P_{\mathbf{x}}(\bar{x}_{m'}) \prod\limits_{i} P_{\mathsf{w}_o}(y_i)$ for any $m' \neq m$, since the codewords generated by the encoder are independent. Therefore,

$$\begin{aligned}
\mathbb{P}\left((\bar{x}_{m'}, \bar{y}) \in \mathsf{T}^{n,\epsilon}_{P_{\mathsf{x}}(x)P_{\mathsf{w}}(y|x)}\right) &= 2^{-nD(P_{\mathsf{x}}(x)P_{\mathsf{w}}(y|x)||P_{\mathsf{x}}(x)P_{\mathsf{w}_o}(y))} \\
&= 2^{-n\sum_{x,y} P_{\mathsf{x}}(x)P_{\mathsf{w}}(y|x)\log\frac{P_{\mathsf{w}}(y|x)}{P_{\mathsf{w}}(y)} + \sum_y P_{\mathsf{w}}(y)\log\frac{P_{\mathsf{w}}(y)}{P_{\mathsf{w}_o}(y)}} \\
&\leq 2^{-nI(W)} \qquad (3.2)
\end{aligned}$$

Due to the strong law of large numbers, $\mathbb{P}\left((\bar{x}_m, \bar{y}) \notin \mathsf{T}^{n,\epsilon}_{P_{\mathsf{x}}(x)P_{\mathsf{w}_o}(y|x)}\right) \to 0$. Hence,

$$\begin{aligned}\mathbb{E}[P_{e,avg}] &\leq \sum_{m' \neq m} \mathbb{P}\left((\bar{x}_{m'}, \bar{y}) \in \bigcup_{W} \mathsf{T}^{n,\epsilon}_{P_{\mathsf{x}}(x)P_{\mathsf{w}}(y|x)}\right) \\ &\leq \sum_{m' \neq m} \sum_{W} \mathbb{P}\left((\bar{x}_{m'}, \bar{y}) \in \mathsf{T}^{n,\epsilon}_{P_{\mathsf{x}}(x)P_{\mathsf{w}}(y|x)}\right) \\ &\leq \mid W \mid 2^{-n\left(\min_{\mathsf{w}} I(W) - R\right)}\end{aligned}$$

where we assumed $M = \lceil 2^{nR} \rceil$.

□

Let $I_\ell(\mathcal{W})$ denote the compound capacity of linear codes. The next theorem shows that linear codes achieves the compound capacity of symmetric channels given in (3.1).

Theorem 3.5. *Let $\mathcal{W}$ be a set of symmetric channels with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, uniform input distribution and transition probabilities $P_{\mathsf{w}}(y \mid x)$ for each $W \in \mathcal{W}$. Then, for any rate $R \geq 0$ such that $R \leq \min_{W \in \mathcal{W}} I(X;Y)$ and any $\epsilon > 0$ there exists a linear code with block decoding error probability $P_e < \epsilon$. As a result,*

$$I_\ell(\mathcal{W}) = \min_{W \in \mathcal{W}} I(W)$$

Proof of Theorem 3.5. From Chapter 6 of [2], we know that the codewords of an affine code, i.e., $x = uG + v$ where v is an arbitrarily fixed sequence in $\mathcal{X}^n$ are pairwise independent. Based on this knowledge, we notice the proof of Theorem 3.2 can be applied to random affine codes. In addition, we recognize that the same performance as an affine code can be obtained with a communication system using a linear code with codewords $x = uG$. Hence, we also expect random linear codes to achieve the symmetric compound capacity $I(\mathcal{W})$. Moreover, since the average error probability can be made arbitrarily small, we know there exists at least one linear code which will have an error probability smaller than or equal to the average error probability. □

3.2 Polar Codes

The compound capacity of polar codes under successive cancellation decoding is analyzed in [6]. The paper provides upper and lower bounds to show that, in general, this capacity does not achieve (3.1). We first explain these existing bounds and the essential idea behind.

We now give a definition of degraded channels in the framework of polar codes.

Definition 3.6. [12]Given two B-DMC W_1 and W_2, assume we apply ℓ recursions of the polarization transformations in (1.5) and (1.6). If W_1 is degraded with respect to W_2, then the Bhattacharyya parameters of all channels satisfy

$$Z(W_1^{(i,2^\ell)}) \leq Z(W_2^{(i,2^\ell)}) \quad \forall i = 1, \dots 2^\ell$$

Let $\mathcal{W}$ be a set of BMS channels and $I_{\mathsf{p}}(\mathcal{W})$ denotes the compound capacity of polar codes under successive cancellation decoding. Then, we can upper bound this capacity as

$$I_{\mathsf{p}}(\mathcal{W}) \leq \min_{W \in \mathcal{W}} I(W)$$

Although the upper bound is trivial from (3.1), we explain why in general we do not expect to have an equality for polar codes. The main reason is related to the construction of polar codes. We know that individual polar codes for each channel W in the set $\mathcal{W}$ can communicate at rates up to $I(W)$ by signaling on the good channel indexes. However, we do not know the intersection of these indexes among different channels in the set, except the particular case in which the channels in the set form a degraded family. Hence, even if each of the individual polar codes can all achieve at least a particular rate R, there might be indexes good for one code that are not good for the others. As a result, none of the codes guarantee to achieve R for all the set of channels. This also emphasizes a link between the compound channel problem and the problem to find universal polarization codes.

Furthermore, we can give as lower bound

$$I_{\mathsf{p}}(\mathcal{W}) \geq \min_{W \in \mathcal{W}} 1 - Z(W) \tag{3.3}$$

The lower bound is not surprising. In fact, the RHS is simply the capacity of the binary erasure channel having Bhattacharyya parameter equals to the smallest one among channels in $\mathcal{W}$. The idea to find an equivalent binary erasure channel W_{BEC} through the equality $Z(W_{\mathsf{BEC}}) = Z(W)$ is a direct consequence of the basic channel transformations

$$\begin{aligned} Z(W^-) &\leq Z(W^-_{\mathsf{BEC}}) \\ Z(W^+) &= Z(W^+_{\mathsf{BEC}}) \end{aligned}$$

and the fact that a binary erasure channel remains a binary erasure channel after the basic channel transformations. Let us introduce the following notation to denote this equivalence

$$\{W\}_{\mathsf{bec}} = W_{\mathsf{BEC}} \quad \text{ such that } \quad Z(W_{\mathsf{BEC}}) = Z(W)$$

Consequently, the channel W is degraded with respect to the equivalent binary erasure channel W_{BEC}, and both channels W and W_{BEC} are degraded with respect to any other binary erasure channel $W_{\mathsf{BEC}'}$ such that $Z(W_{\mathsf{BEC}}) \leq Z(W_{\mathsf{BEC}'})$. Moreover, this lower bound suggests an explicit code construction method for compound channels: design the code to work with the binary erasure channel W_{BEC} such that

$$W_{\mathsf{BEC}} = \{W\}_{\mathsf{bec}}$$

where

$$W \in \mathcal{W} : Z(W) = \max_{W' \in \mathcal{W}} Z(W')$$

This code will guarantee the same performance in all the channels in $\mathcal{W}$.

These two bounds were significantly improved in [6] to give

$$I_{\mathsf{p}}(\mathcal{W}) \leq \frac{1}{2^\ell} \sum_i \min_{W \in \mathcal{W}} I(W^{(i)})$$
$$I_{\mathsf{p}}(\mathcal{W}) \geq \frac{1}{2^\ell} \sum_i \min_{W \in \mathcal{W}} \{1 - I(W^{(i)})\}$$

where $i = 1, \ldots, 2^\ell$ and both bounds are monotone in $\ell \in \mathbb{N}$.

Chapter 4

Main Results

In this chapter, we generalize results on the extremality of the binary erasure channel and the binary symmetric channel with respect to the parameters $I(W)$ and $Z(W)$ to $E_0(\rho, W)$, when the basic channel polarization transformations are applied. Then, as special cases, we derive expressions for the binary erasure channel and the binary symmetric channel. Then, we describe $E_0(\rho, W^-)$ and $E_0(\rho, W^+)$ in terms of Rényi's entropy functions. Based on this representation, we conjecture an inequality between $E_0(\rho, W)$, $E_0(\rho, W^-)$, and $E_0(\rho, W^+)$. Finally, we conclude this chapter by the section "More Extremalities", where we examine whether the result obtained in [3] can be extended. The authors show that the binary erasure channel and the binary symmetric channel are extremal with respect to $R(\rho, W)$ and $E_0(\rho, W)$. Following the idea, an interesting question we raise is: Can we derive similar extremal properties for $E_0(\rho_1, W)$ and $R(\rho_2, W)$, i.e for the cases we operate at different ρ values?

4.1 Extremality of the Basic Channel Transformations

We start with a lemma proved in [3]. Then using the same technique, we show that similar results exist for $E_0(\rho, W^-)$ and $E_0(\rho, W^+)$ in Lemma 4.3 and Lemma 4.4, respectively.

Lemma 4.1. *Given a channel W and $\rho \in [0, 1]$, there exist a random variable Z taking values in the $[0, 1]$ interval such that*

$$E_0(\rho, W) = -\log \mathbb{E}\left[g(\rho, Z)\right] \tag{4.1}$$

where

$$g(\rho, z) = \left(\frac{1}{2}(1+z)^{\frac{1}{1+\rho}} + \frac{1}{2}(1-z)^{\frac{1}{1+\rho}}\right)^{1+\rho}$$

Proof of Lemma 4.1.

$$E_0(\rho, W) = -\log \sum_y \left[\ \frac{1}{2}P(y \mid 0)^{\frac{1}{1+\rho}} + \frac{1}{2}P(y \mid 1)^{\frac{1}{1+\rho}}\ \right]^{1+\rho}$$

We define,

$$q(y) = \frac{P(y \mid 0) + P(y \mid 1)}{2} \qquad \Delta(y) = \frac{P(y \mid 0) - P(y \mid 1)}{P(y \mid 0) + P(y \mid 1)} \tag{4.2}$$

$$\Rightarrow \quad \frac{P(y \mid 0)}{q(y)} = 1 + \Delta(y) \qquad \frac{p(y \mid 1)}{q(y)} = 1 - \Delta(y)$$

Then, one can define a random variable $Z = |\Delta(Y)| \in [0, 1]$ where Y has the probability distribution $q(y)$, and obtain (4.1) by simple manipulations. □

Lemma 4.2. *Given a channel W and $\rho \in [0, 1]$, there exist independent random variables Z_1 and Z_2 taking values in the $[0, 1]$ interval such that*

$$E_0(\rho, W^-) = -\log \mathbb{E}\left[g(\rho, Z_1 Z_2)\right] \tag{4.3}$$

where $g(\rho, z)$ is given by (4.2).

Proof of Lemma 4.2. From the definition of channel W^- in (1.5), we can write

$$\begin{aligned}
E_0(\rho, W^-) &= -\log \sum_{y_1, y_2} \left[\frac{1}{2} P_{W^-}(y_1, y_2 \mid 0)^{\frac{1}{1+\rho}} + \frac{1}{2} P_{W^-}(y_1, y_2 \mid 1)^{\frac{1}{1+\rho}} \right]^{1+\rho} \\
&= -\log \sum_{y_1, y_2} \Bigg[\frac{1}{2} \left(\frac{1}{2} P(y_1 \mid 0) P(y_2 \mid 0) + \frac{1}{2} P(y_1 \mid 1) P(y_2 \mid 1) \right)^{\frac{1}{1+\rho}} \\
&\qquad + \frac{1}{2} \left(\frac{1}{2} P(y_1 \mid 1) P(y_2 \mid 0) + \frac{1}{2} P(y_1 \mid 0) P(y_2 \mid 1) \right)^{\frac{1}{1+\rho}} \Bigg]^{1+\rho} \\
&= -\log \sum_{y_1 y_2} \Bigg[\frac{1}{2} \left(\frac{1}{2} \right)^{\frac{1}{1+\rho}} q\left(y_1\right)^{\frac{1}{1+\rho}} q\left(y_2\right)^{\frac{1}{1+\rho}} \\
&\qquad \left(\left(1 + \Delta\left(y_1\right)\right) \left(1 + \Delta\left(y_2\right)\right) + \left(1 - \Delta\left(y_1\right)\right) \left(1 - \Delta\left(y_2\right)\right) \right)^{\frac{1}{1+\rho}} \\
&\qquad + \left(\left(1 - \Delta\left(y_1\right)\right) \left(1 + \Delta\left(y_2\right)\right) + \left(1 + \Delta\left(y_1\right)\right) \left(1 - \Delta\left(y_2\right)\right) \right)^{\frac{1}{1+\rho}} \Bigg]^{1+\rho} \\
&= -\log \sum_{y_1 y_2} q(y_1)\ q(y_2) \left[\frac{1}{2} \left(1 + \Delta(y_1)\Delta(y_2)\right)^{\frac{1}{1+\rho}} + \frac{1}{2} \left(1 - \Delta(y_1)\Delta(y_2)\right)^{\frac{1}{1+\rho}} \right]^{1+\rho}
\end{aligned}$$

where we used (4.2). We can now define two independent random variables $Z_1 = |\Delta(Y_1)|$ and $Z_2 = |\Delta(Y_2)|$ where $Y_1 \sim q(y_1)$ and $Y_2 \sim q(y_2)$. Moreover, we note that by definition both Z_1 and Z_2 takes values in the $[0, 1]$ interval. From this construction, the lemma follows. □

Lemma 4.3. *Given a channel W and $\rho \in [0, 1]$, there exist independent random variables Z_1 and Z_2 taking values in the $[0, 1]$ interval such that*

$$E_0(\rho, W^+) = -\log \mathbb{E}\left[\frac{1}{2}(1 + Z_1 Z_2)\ g(\rho, \frac{Z_1 + Z_2}{1 + Z_1 Z_2}) + \frac{1}{2}(1 - Z_1 Z_2)\ g(\rho, \frac{Z_1 - Z_2}{1 - Z_1 Z_2}) \right] \tag{4.4}$$

where $g(\rho, z)$ is given by (4.2).

Proof of Lemma 4.3. From the definition of channel W^+ in (1.6), we can write

$$
\begin{aligned}
&E_0(\rho, W^+) \\
&= -\log \sum_{y_1,y_2,u} \left[\frac{1}{2} P_{\mathsf{W}^+}(y_1, y_2, u \mid 0)^{\frac{1}{1+\rho}} + \frac{1}{2} P_{\mathsf{W}^+}(y_1, y_2, u \mid 1)^{\frac{1}{1+\rho}} \right]^{1+\rho} \\
&= -\log \sum_{y_1,y_2,u} \left[\frac{1}{2} \left(\frac{1}{2} P(y_1 \mid u) P(y_2 \mid 0) \right)^{\frac{1}{1+\rho}} + \frac{1}{2} \left(\frac{1}{2} P(y_1 \mid u \oplus 1) P(y_2 \mid 1) \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \\
&= -\log \sum_{y_1,y_2} \Bigg(\left[\frac{1}{2} \left(\frac{1}{2} P(y_1 \mid 0) P(y_2 \mid 0) \right)^{\frac{1}{1+\rho}} + \frac{1}{2} \left(\frac{1}{2} P(y_1 \mid 1) P(y_2 \mid 1) \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \\
&\qquad + \left[\frac{1}{2} \left(\frac{1}{2} P(y_1 \mid 1) p(y_2 \mid 0) \right)^{\frac{1}{1+\rho}} + \frac{1}{2} \left(\frac{1}{2} P(y_1 \mid 0) p(y_2 \mid 1) \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \Bigg)
\end{aligned}
$$

Using (4.2), we have

$$
\begin{aligned}
&E_0(\rho, W^+) \\
&= -\log \sum_{y_1 y_2} \frac{1}{2} \; q(y_1) \; q(y_2) \\
&\qquad \Bigg(\left[\left((1 + \Delta(y_1)) (1 + \Delta(y_2)) \right)^{\frac{1}{1+\rho}} + \left((1 - \Delta(y_1)) (1 - \Delta(y_2)) \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \\
&\qquad + \left[\left((1 - \Delta(y_1)) (1 + \Delta(y_2)) \right)^{\frac{1}{1+\rho}} + \left((1 - \Delta(y_1)) (1 + \Delta(y_2)) \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \Bigg) \\
&= -\log \Bigg(\sum_{y_1 y_2} \frac{1}{2} \; q(y_1) \; q(y_2) \; (1 + \Delta(y_1)\Delta(y_2)) \\
&\qquad \left[\frac{1}{2} \left(1 + \frac{\Delta(y_1) + \Delta(y_2)}{1 + \Delta(y_1)\Delta(y_2)} \right)^{\frac{1}{1+\rho}} + \frac{1}{2} \left(1 - \frac{\Delta(y_1) + \Delta(y_2)}{1 + \Delta(y_1)\Delta(y_2)} \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \\
&\qquad + \sum_{y_1 y_2} \frac{1}{2} \; q(y_1) \; q(y_2) \; (1 - \Delta(y_1)\Delta(y_2)) \\
&\qquad \left[\frac{1}{2} \left(1 + \frac{\Delta(y_1) - \Delta(y_2)}{1 - \Delta(y_1)\Delta(y_2)} \right)^{\frac{1}{1+\rho}} + \frac{1}{2} \left(1 - \frac{\Delta(y_1) - \Delta(y_2)}{1 - \Delta(y_1)\Delta(y_2)} \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \Bigg) \\
&= -\log \Bigg(\sum_{y_1 y_2} \frac{1}{2} \; q(y_1) \; q(y_2) \; (1 + \Delta(y_1)\Delta(y_2)) \; g\left(\rho, \frac{\Delta(y_1) + \Delta(y_2)}{1 + \Delta(y_1)\Delta(y_2)} \right) \\
&\qquad + \sum_{y_1 y_2} \frac{1}{2} \; q(y_1) \; q(y_2) \; (1 - \Delta(y_1)\Delta(y_2)) \; g\left(\rho, \frac{\Delta(y_1) - \Delta(y_2)}{1 - \Delta(y_1)\Delta(y_2)} \right) \Bigg)
\end{aligned}
$$

where $g(\rho, z)$ is defined in (4.2).

Similar to the $E_0(\rho, W^-)$ case, we want to define two independent random variables $Z_1 = |\Delta(Y_1)|$ and $Z_2 = |\Delta(Y_2)|$ where $Y_1 \sim q(y_1)$ and $Y_2 \sim q(y_2)$ and both

Z_1 and Z_2 takes values in the $[0,1]$ interval. However, we should first check whether this construction is equivalent to the above equation. We note that $\Delta(y) \in [-1,1]$. When both $\Delta(y_1)$ and $\Delta(y_2)$ are positive, i.e $\in [0,1]$, we can easily see that

$$(1+\Delta(y_1)\Delta(y_2))\;\; g(\rho, \frac{\Delta(y_1)+\Delta(y_2)}{1+\Delta(y_1)\Delta(y_2)}) = (1+Z_1Z_2)\;\; g(\rho, \frac{Z_1+Z_2}{1+Z_1Z_2})$$
$$(1-\Delta(y_1)\Delta(y_2))\;\; g(\rho, \frac{\Delta(y_1)-\Delta(y_2)}{1-\Delta(y_1)\Delta(y_2)}) = (1-Z_1Z_2)\;\; g(\rho, \frac{Z_1-Z_2}{1-Z_1Z_2})$$

When $\Delta(y_1) \in [0,1]$ and $\Delta(y_2) \in [-1,0)$, we note that

$$(1+\Delta(y_1)\Delta(y_2))\;\; g(\rho, \frac{\Delta(y_1)+\Delta(y_2)}{1+\Delta(y_1)\Delta(y_2)}) = (1-Z_1Z_2)\;\; g(\rho, \frac{Z_1-Z_2}{1-Z_1Z_2})$$
$$(1-\Delta(y_1)\Delta(y_2))\;\; g(\rho, \frac{\Delta(y_1)-\Delta(y_2)}{1-\Delta(y_1)\Delta(y_2)}) = (1+Z_1Z_2)\;\; g(\rho, \frac{Z_1+Z_2}{1+Z_1Z_2})$$

Since we are interested in the sum of the above two parts, we can see that the construction we propose is still equivalent. Finally, for the other possible cases of $\Delta(y_1)$ and $\Delta(y_2)$ values, similar results hold since the function $g(\rho, z)$ is symmetric with respect to $z = 0$. This concludes the proof. □

Before the extremality analysis, we make an important remark.

Remark 4.4. The random variable Z_{BEC} of a binary erasure channel is $\{0,1\}$ valued. The random variable Z_{BSC} of a binary symmetric channel is constant z_{BSC}.

Theorem 4.5. *Given a fixed $\rho \in (0,1)$ and a channel W, a binary erasure channel W_{BEC}, and a binary symmetric channel W_{BSC} such that the following inequality holds*

$$E_0(\rho, W_{\mathsf{BEC}}) = E_0(\rho, W_{\mathsf{BSC}}) = E_0(\rho, W)$$

Then,

$$E_0(\rho, W^-_{\mathsf{BEC}}) \leq E_0(\rho, W^-) \leq E_0(\rho, W^-_{\mathsf{BSC}}) \tag{4.5}$$

Proof of Theorem 4.5. By lemmas 4.2 and 4.3, we have

$$\exp\{-E_0(\rho, W)\} = \mathbb{E}\left[g(\rho, Z)\right] \quad \text{and} \quad \exp\{-E_0(\rho, W)\} = \mathbb{E}[g(\rho, Z_1Z_2)]$$

where Z_1 and Z_2 are independent random variables. Moreover by Remark 4.4, we know $Z_{\mathsf{BSC}} = z_{\mathsf{BSC}}$ and $Z_{\mathsf{BEC}} \in \{0,1\}$. Hence,

$$\exp\{-E_0(\rho, W^-_{\mathsf{BSC}})\} = g(\rho, z_{\mathsf{BSC}} z_{\mathsf{BSC}})$$

and

$$\begin{aligned}
&\exp\{-E_0(\rho, W^-_{\mathsf{BEC}})\} \\
&= P(Z_{\mathsf{BEC}}=0)^2 g(\rho,0) + 2P(Z_{\mathsf{BEC}}=0)P(Z_{\mathsf{BEC}}=1)g(\rho,0) + P(Z_{\mathsf{BEC}}=1)^2 g(\rho,1) \\
&= P(Z_{\mathsf{BEC}}=0)^2 + 2P(Z_{\mathsf{BEC}}=0)\left(1 - P(Z_{\mathsf{BEC}}=0)\right) + \left(1 - P(Z_{\mathsf{BEC}}=0)^2\right) 2^{-\rho} \\
&= \left[2P(Z_{\mathsf{BEC}}=0) - P(Z_{\mathsf{BEC}}=0)^2\right](1-2^{-\rho}) + 2^{-\rho}
\end{aligned}$$

We define the function $F_{z,\rho}(t) : [2^{-\rho}, 1] \to [g(\rho, z), g(\rho, 0)]$ as

$$F_{z,\rho}(t) = g(\rho, zg^{-1}(\rho, t)) \tag{4.6}$$

We prove in Appendix B that for fixed values of ρ and z, $F_{z,\rho}(t)$ is convex with respect to t.

Given $E_0(\rho, W) = E_0(\rho, W_{\mathsf{BSC}})$, we know

$$\mathbb{E}\left[g(\rho, Z)\right] = g(\rho, z_{\mathsf{BSC}})$$

Then, using Jensen's inequality we obtain

$$\begin{aligned}
\exp\{-E_0(\rho, W^-)\} &= \mathbb{E}_{Z_1}[\mathbb{E}_{Z_2}[F_{z_1,\rho}\left(g(\rho, Z_2)\right) \mid Z_1 = z_1]] \\
&\geq \mathbb{E}_{Z_1}[F_{Z_1,\rho}\left(\mathbb{E}_{Z_2}[g(\rho, Z_2)]\right)] \\
&= \mathbb{E}_{Z_1}[F_{Z_1,\rho}\left(g(\rho, z_{\mathsf{BSC}})\right)] \\
&\overset{(1)}{=} \mathbb{E}_{Z_1}[F_{z_{\mathsf{BSC}},\rho}\left(g(\rho, Z_1)\right)] \\
&\geq F_{z_{\mathsf{BSC}},\rho}\left(\mathbb{E}_{Z_1}[g(\rho, Z_1)]\right) \\
&= F_{z_{\mathsf{BSC}},\rho}\left(g(\rho, z_{\mathsf{BSC}})\right) \\
&= \exp\{-E_0(\rho, W^-_{\mathsf{BSC}})\}
\end{aligned}$$

where (1) follows by symmetry of the variables z_1 and z_2. Similarly, given $E_0(\rho, W) = E_0(\rho, W_{BEC})$, we have

$$\begin{aligned}
\mathbb{E}\left[g(\rho, Z)\right] = \mathbb{E}\left[g(\rho, Z_{\mathsf{BEC}})\right] &= P(Z_{\mathsf{BEC}} = 0)g(\rho, 0) + P(Z_{\mathsf{BEC}} = 1)g(\rho, 1) \\
&= P(Z_{\mathsf{BEC}} = 0)(1 - 2^{-\rho}) + 2^{-\rho}
\end{aligned}$$

Due to convexity, we also have

$$F_{z,\rho}(t) \leq g(\rho, 0) + \frac{g(\rho, z_1) - g(\rho, 0)}{2^{-\rho} - 1}(t - 1) = 1 + \frac{g(\rho, z_1) - 1}{2^{-\rho} - 1}(t - 1)$$

Therefore,

$$\begin{aligned}
\exp\{-E_0(\rho, W^-)\} =& \mathbb{E}_{Z_1}[\mathbb{E}_{Z_2}[F_{z_1,\rho}\left(g(\rho, Z_2)\right) \mid Z_1 = z_1]] \\
&\leq \mathbb{E}_{Z_1}\left[1 + \frac{g(\rho, Z_1) - 1}{2^{-\rho} - 1}(\mathbb{E}_{Z_2}[g(\rho, Z_2)] - 1)\right] \\
&= 1 + \frac{\mathbb{E}_{Z_1}[g(\rho, Z_1)] - 1}{2^{-\rho} - 1}(\mathbb{E}_{Z_2}[g(\rho, Z_2)] - 1) \\
&= \frac{2^{-\rho} - 1 + \left(P(Z_{\mathsf{BEC}} = 0)(1 - 2^{-\rho}) + 2^{-\rho} - 1\right)^2}{2^{-\rho} - 1} \\
&= \frac{(2^{-\rho} - 1)\left(1 + (1 - P(Z_{\mathsf{BEC}} = 0))^2\left(2^{-\rho} - 1\right)\right)}{2^{-\rho} - 1} \\
&= \left[2P(Z_{\mathsf{BEC}} = 0) - P(Z_{\mathsf{BEC}} = 0)^2\right](1 - 2^{-\rho}) + 2^{-\rho} \\
&= \exp\{-E_0(\rho, W^-_{\mathsf{BEC}})\}
\end{aligned}$$

□

In Theorem 4.5, we showed that among all B-DMC's W of fixed $E_0(\rho, W)$, the binary erasure channel W^- transformation results in a lower bound to any $E_0(\rho, W^-)$ and the binary symmetric channel's one in an upper bound to any $E_0(\rho, W^-)$. In the next theorem, we state a similar result for the W^+ transformation, except the difference that the binary erasure channel W^+ transformation provides an upper bound and the binary symmetric channel's one a lower bound to any $E_0(\rho, W^+)$ in this case.

Theorem 4.6. *Given a fixed $\rho \in (0, 1)$ and a channel W, a binary erasure channel W_{BEC}, and a binary symmetric channel W_{BSC} such that the following inequality holds*

$$E_0(\rho, W_{\mathsf{BEC}}) = E_0(\rho, W_{\mathsf{BSC}}) = E_0(\rho, W)$$

Then,

$$E_0(\rho, W^+_{\mathsf{BSC}}) \leq E_0(\rho, W^+) \leq E_0(\rho, W^+_{\mathsf{BEC}}) \tag{4.7}$$

Proof of Theorem 4.6. By Lemmas 4.2 and 4.3, we can write

$$\exp\{-E_0(\rho, W)\} = \mathbb{E}\left[g(\rho, Z)\right]$$
$$\exp\{-E_0(\rho, W^+)\} = \mathbb{E}\left[\ \frac{1}{2}(1 + Z_1 Z_2)\ \ g(\rho, \frac{Z_1 + Z_2}{1 + Z_1 Z_2}) + \frac{1}{2}(1 - Z_1 Z_2)\ \ g(\rho, \frac{Z_1 - Z_2}{1 - Z_1 Z_2})\ \right]$$

where Z_1 and Z_2 are independent random variables. Moreover by Remark 4.4, we know $Z_{\mathsf{BSC}} = z_{\mathsf{BSC}}$ and $Z_{\mathsf{BEC}} \in \{0, 1\}$. Hence,

$$\exp\{-E_0(\rho, W^+_{\mathsf{BSC}})\} = \frac{1}{2}(1 + z^2_{\mathsf{BSC}})\ \ g(\rho, \frac{2z_{\mathsf{BSC}}}{1 + z^2_{\mathsf{BSC}}}) + \frac{1}{2}(1 - z^2_{\mathsf{BSC}})$$

and

$$\begin{aligned}
&\exp\{-E_0(\rho, W^+_{\mathsf{BEC}})\} \\
&= P(Z_{\mathsf{BEC}} = 0)^2 g(\rho, 0) + 2P(Z_{\mathsf{BEC}} = 0)P(Z_{\mathsf{BEC}} = 1)g(\rho, 1) + P(Z_{\mathsf{BEC}} = 1)^2 g(\rho, 1) \\
&= (1 - P(Z_{\mathsf{BEC}} = 1))^2 + 2(1 - P(Z_{\mathsf{BEC}} = 1))P(Z_{\mathsf{BEC}} = 1)2^{-\rho} + P(Z_{\mathsf{BEC}} = 1)^2 2^{-\rho} \\
&= \left(1 - P(Z_{\mathsf{BEC}} = 1)\right)^2 \left(1 - 2^{-\rho}\right) + 2^{-\rho}.
\end{aligned}$$

We define the function

$$t_3(z_1, z_2) = \frac{1}{2}(1 + z_1 z_2)\ \ g(\rho, \frac{z_1 + z_2}{1 + z_1 z_2}) + \frac{1}{2}(1 - z_1 z_2)\ \ g(\rho, \frac{z_1 - z_2}{1 - z_1 z_2}) \tag{4.8}$$

where $\rho, z_1, z_2 \in [0, 1]$. Then, we have

$$\mathbb{E}\left[t_3(Z_1, Z_2)\right] = \exp\{-E_0(\rho, W^+)\}$$

We show in Appendix C that $t_3(z_1, z_2)$ is a concave decreasing function in both z_1 and z_2 separately. Note also that $t_3(z_1, z_2)$ is symmetric in the variables z_1, and z_2.

Let the function $H_{z,\rho}(t) : [2^{-\rho}, 1] \to [2^{-\rho}, g(\rho, z)]$ be defined as

$$H_{z,\rho}(t) = t_3(g^{-1}(\rho, t), z) \tag{4.9}$$

where both $z, \rho \in [0,1]$. For fixed values of ρ and z, the function $H_{z,\rho}(t)$ is concave with respect to the variable t. We refer the readers to Appendix F for a proof.

The proof of the theorem can be completed following similar steps to the W^- case. We define $T_1 = g(\rho, Z_1)$, and $T_2 = g(\rho, Z_2)$. Then, using the concavity of the function $H_{z,\rho}(t)$ with respect to t, and symmetry of Z_1 and Z_2, we obtain

$$\begin{aligned}\exp\{-E_0(\rho, W^+)\} &= \mathbb{E}\left[H_{g^{-1}(\rho,T_2),\rho}(T_1)\right] \\ &\leq h_\rho(z_{\mathsf{BSC}}, z_{\mathsf{BSC}}) = \exp\{-E_0(\rho, W^+_{\mathsf{BSC}})\}\end{aligned}$$

and

$$\begin{aligned}\exp\{-E_0(\rho, W^+)\} &= \mathbb{E}\left[H_{g^{-1}(\rho,T_2),\rho}(T_1)\right] \\ &\geq 2^{-\rho} + P(Z_{\mathsf{BEC}} = 0)^2\left(1 - 2^{-\rho}\right) = \exp\{-E_0(\rho, W^+_{\mathsf{BEC}})\}.\end{aligned}$$

□

Theorem 4.5 and Theorem 4.6 show that the binary erasure and binary symmetric channels appear on reversed sides of the inequalities for $E_0(\rho, W^-)$ and $E_0(\rho, W^+)$. On the other hand, recall that the idea behind the lower bound in (3.3) for the compound capacity of polar codes relies on finding an equivalent binary erasure channels as an upper bound to the process Z_ℓ. For that reason, we re-arrange the inequalities in Proposition 4.7 such that the quantities for the binary erasure channel and the binary symmetric channel shift sides at the cost of having less tight bounds.

Proposition 4.7. *Given a fixed $\rho \in (0,1)$ and a channel W, a binary erasure channel W_{BEC}, and a binary symmetric channel W_{BSC} such that the following inequality holds*

$$E_0(\rho, W_{\mathsf{BEC}}) = E_0(\rho, W_{\mathsf{BSC}}) = E_0(\rho, W)$$

Then, W^+_{BEC} and W^+_{BSC} correspond to extreme values of W^+ transformation with

$$\rho E_0(\rho, W^+_{\mathsf{BEC}}) \leq E_0(\rho, W^+) \leq \frac{1}{\rho} E_0(\rho, W^+_{\mathsf{BSC}}) \tag{4.10}$$

Proof of Proposition 4.7. To prove the left inequality in (4.10), we show that the following inequality holds

$$\rho E_0(\rho, W^+_{\mathsf{BEC}}) \leq E_0(\rho, W^+_{\mathsf{BSC}}) \tag{4.11}$$

Let's first note that when $E_0(\rho, W_{\mathsf{BEC}}) = E_0(\rho, W_{\mathsf{BSC}})$

$$\begin{aligned}\rho E_0(\rho, W^+_{\mathsf{BEC}})\Big|_{\rho=1} &= E_0(\rho, W^+_{\mathsf{BSC}})\Big|_{\rho=1} \\ \rho E_0(\rho, W^+_{\mathsf{BEC}})\Big|_{\rho=0} &= E_0(\rho, W^+_{\mathsf{BSC}})\Big|_{\rho=0}\end{aligned}$$

We already know that E_0 is a concave increasing function in ρ. In addition, we next show that $\rho E_0(\rho, W^+_{\mathsf{BEC}})$ is a convex function in ρ. Let $\epsilon_{\mathsf{bec}} \in [0,1]$ denotes the channel parameter of W_{BEC}, then

$$\frac{\partial^2}{\partial \rho^2} \rho E_0(\rho, W^+_{\mathsf{BEC}}) = \frac{\left(1 - \epsilon^2_{\mathsf{bec}}\right)\left(2 - \epsilon^2_{\mathsf{bec}}\left(-2\left(2^\rho - 1\right) + 2^\rho \rho \log 2\right)\right)}{\left(1 + \left(2^\rho - 1\right)\epsilon^2_{\mathsf{bec}}\right)^2} \geq 0$$

since

$$\frac{\partial}{\partial \rho}\left(-2\left(2^{\rho}-1\right)+2^{\rho}\rho\log 2\right)=-2^{\rho}\log 2\left(1-g\log 2\right)\leq 0$$
$$\Rightarrow\left(-2\left(2^{\rho}-1\right)+2^{\rho}\rho\log 2\right)\leq 0$$

As a result, we conclude (4.11) holds.

On the other hand, we already mentioned that $E_0(\rho,W)/\rho$ is a decreasing function in ρ in Chapter 2. Moreover, we note that

$$E_0(\rho,W_{\mathsf{BEC}}^{+})\Big|_{\rho=1}=\frac{1}{\rho}E_0(\rho,W_{\mathsf{BSC}}^{+})\Big|_{\rho=1}$$

Therefore,

$$E_0(\rho,W_{\mathsf{BEC}}^{+})\leq\frac{1}{\rho}E_0(\rho,W_{\mathsf{BSC}}^{+})$$

which proves the right inequality in (4.10). □

4.2 Analysis of Special Cases: BEC and BSC

In the previous section, we showed that the basic channel transformations satisfy extremal properties by the binary erasure channel and the binary symmetric channel. For that reason, we provide the expressions for these channels in terms of their channel parameters. We also approve the fact that, for a binary erasure channel, the channels we get after applying the basic channel transformations are also binary erasure channels. On the other hand, for a binary symmetric channel, while the channel we get after the W^{-} transformation is a binary symmetric channel, the one we have after the W^{+} transformation is not. Since the derivations are quite simple, we provide directly the results.

Let W_{BSC} be a binary symmetric channel with crossover probability ϵ_{bsc}. Then,

$$E_0(\rho,W_{\mathsf{BSC}})=-\log 2^{-\rho}\left(\left(1-\epsilon_{\mathsf{bsc}}\right)^{\frac{1}{1+\rho}}+\left(\epsilon_{\mathsf{bsc}}\right)^{\frac{1}{1+\rho}}\right)^{1+\rho}$$
$$E_0(\rho,W_{\mathsf{BSC}}^{-})=-\log 2^{-\rho}\left(\left(1-\epsilon_{\mathsf{bsc}}^{-}\right)^{\frac{1}{1+\rho}}+\left(\epsilon_{\mathsf{bsc}}^{-}\right)^{\frac{1}{1+\rho}}\right)^{1+\rho}\quad\text{with }\ \epsilon_{bsc}^{-}=2\epsilon_{\mathsf{bsc}}-2\epsilon_{\mathsf{bsc}}^{2}$$
$$E_0(\rho,W_{\mathsf{BSC}}^{+})=-\log\left(2^{-\rho}\left(\left(1-\epsilon_{\mathsf{bsc}}\right)^{\frac{2}{1+\rho}}+\left(\epsilon_{\mathsf{bsc}}\right)^{\frac{2}{1+\rho}}\right)^{1+\rho}+2\epsilon_{\mathsf{bsc}}\left(1-\epsilon_{\mathsf{bsc}}\right)\right)$$

Let W_{BEC} be a binary erasure channel with erasure probability ϵ_{bec}. Then,

$$E_0(\rho,W_{\mathsf{BEC}})=-\log\left(2^{-\rho}\left(1-\epsilon_{\mathsf{bec}}\right)+\epsilon_{\mathsf{bec}}\right)$$
$$E_0(\rho,W_{\mathsf{BEC}}^{-})=-\log\left(2^{-\rho}\left(1-\epsilon_{\mathsf{bec}}^{-}\right)+\epsilon_{\mathsf{bec}}^{-}\right)\quad\text{with }\ \epsilon_{\mathsf{bec}}^{-}=2\epsilon_{\mathsf{bec}}-\epsilon_{\mathsf{bec}}^{2}$$
$$E_0(\rho,W_{\mathsf{BEC}}^{+})=-\log\left(2^{-\rho}\left(1-\epsilon_{\mathsf{bec}}^{+}\right)+\epsilon_{\mathsf{bec}}^{+}\right)\quad\text{with }\ \epsilon_{\mathsf{bec}}^{+}=\epsilon_{\mathsf{bec}}^{2}$$

From $E_0(\rho, W_{\mathsf{BEC}})$ expression, we have

$$\epsilon_{\mathsf{bec}} = \frac{2^{\rho} 2^{-E_0(\rho, W_{\mathsf{BEC}})} - 1}{2^{\rho} - 1}$$

Moreover, we know that the Bhattacharyya parameter of a binary erasure channel satisfies $Z(W_{\mathsf{BEC}}) = \epsilon_{\mathsf{bec}}$. This parameter provides tighter bounds than $E_0(1, W)$ in [1], and is used in the subsequent analysis. This gives the idea to define a similar quantity to $Z(W)$, referred as $Z(\rho, W)$, which reflects the dependence on the value of ρ.

$$Z(\rho, W) = \frac{2^{\rho} 2^{-E_0(\rho, W)} - 1}{2^{\rho} - 1}$$

Using the results we derived in the previous section, we now investigate how $Z(\rho, W)$ is affected by the basic channel transformations.

Proposition 4.8. *Given a channel W, a binary erasure channel W_{BEC}, and a binary symmetric channel W_{BSC}, such that the following inequality holds*

$$Z(\rho, W) = Z(\rho, W_{\mathsf{BEC}}) = Z(\rho, W_{\mathsf{BSC}})$$

Then, we have

$$\begin{aligned} Z(\rho, W_{\mathsf{BSC}}^{-}) \;&\leq\; Z(\rho, W^{-}) \;\leq\; Z(\rho, W_{\mathsf{BEC}}^{-}) = 2Z(\rho, W_{\mathsf{BEC}}) - Z(\rho, W_{BEC})^2 \\ Z(\rho, W_{\mathsf{BEC}})^2 = Z(\rho, W_{\mathsf{BEC}}^{+}) \;&\leq\; Z(\rho, W^{+}) \;\leq\; Z(\rho, W_{\mathsf{BSC}}^{+}) \end{aligned}$$

Proof of Proposition 4.8. The above inequalities are direct consequences of Theorem 4.5 and Theorem 4.6, respectively. □

4.3 The Basic Channel Transformations and Rényi's Entropy Description

In Section 2.2 of Chapter 1, the function $E_0(\rho, W)/\rho$ is defined in terms of Rényi's entropy functions. Similarly, we can characterize $E_0(\rho, W^-)/\rho$ and $E_0(\rho, W^+)/\rho$ in terms of Rényi's entropy functions. Based on these definitions, we conjecture an inequality between $E_0(\rho, W)$, $E_0(\rho, W^-)$, and $E_0(\rho, W^+)$. Given a DMC W with uniform input distribution, consider the W^- and W^+ transformations defined in (1.5) and (1.6), respectively. We showed the resulting channel configuration in Figure 1.2. Using (2.7), $E_0(\rho, W^-)$ and $E_0(\rho, W^+)$ can be defined as follows

$$\begin{aligned}\frac{E_0(\rho, W^-)}{\rho} &= H_{\frac{1}{1+\rho}}(U_1) - H_{\frac{1}{1+\rho}}(U_1 \mid Y_1Y_2)\\ \frac{E_0(\rho, W_+)}{\rho} &= H_{\frac{1}{1+\rho}}(U_2) - H_{\frac{1}{1+\rho}}(U_2 \mid Y_1Y_2U_1)\end{aligned}$$

In addition, as we use two independent copies of W, we can write

$$\begin{aligned}2\frac{E_0(\rho, W)}{\rho} &= -\frac{1}{\rho}\log\left(\sum_{y_1}\left(\sum_{x_1} p(x_1)p(y_1 \mid x_1)^{\frac{1}{1+\rho}}\right)^{1+\rho}\sum_{y_2}\left(\sum_{x_2} p(x_2)p(y_2 \mid x_2)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right)\\ &= \log\left(\sum_{y_1y_2}\left(\sum_{x_1x_2} p(x_1)p(x_2)p(y_1 \mid x_1)^{\frac{1}{1+\rho}}p(y_2 \mid x_2)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right)\\ &= \log\left(\sum_{y_1y_2}\left(\sum_{x_1x_2} p(x_1, x_2)p(y_1y_2 \mid x_1x_2)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right)\\ &= H_{\frac{1}{1+\rho}}(X_1X_2) - H_{\frac{1}{1+\rho}}(X_1X_2 \mid Y_1Y_2)\end{aligned}$$

Since the mapping between $(x_1, x_2) \to (u_1, u_2)$ is one-to-one, this is equivalent to

$$2\frac{E_0(\rho, W)}{\rho} = H_{\frac{1}{1+\rho}}(U_1U_2) - H_{\frac{1}{1+\rho}}(U_1U_2 \mid Y_1Y_2)$$

We now inquire whether a particular relation can be derived between $E_0(\rho, W)$, $E_0(\rho, W^-)$, and $E_0(\rho, W^+)$ using the above definitions. For that purpose, we observe that such a relation, if it exists, would be an implication of a relation between $H_{\frac{1}{1+\rho}}(U_1 \mid Y_1Y_2)$, $H_{\frac{1}{1+\rho}}(U_2 \mid Y_1Y_2U_1)$, and $H_{\frac{1}{1+\rho}}(U_1U_2 \mid Y_1Y_2)$. We next conjecture our observation regarding to this relation. Although we could not end up with a proof, neither we found a counter example.

Conjecture 4.9. The channels W, W^-, and W^+ satisfy the following relationship

$$\frac{1}{2}\left(E_0(\rho, W^-) + E_0(\rho, W^+)\right) \geq E_0(\rho, W)$$

4.4 More Extremalities

In chapter 2, we mentioned that the binary erasure channel and the binary symmetric channel are extremal with respect to $R(\rho, W)$ and $E_0(\rho, W)$. In this section, we ask whether we can derive similar extremal properties for $E_0(\rho_1, W)$ and $R(\rho_2, W)$, i.e., for the cases we operate at different ρ values. To have an idea about the answer, we first carried a numerical experiment.

4.4.1 Numerical Experiment Results

We briefly explain the experiment and discuss results. We generated binary input channels for different size of the output alphabet, and calculated the $E_0(\rho_1, W)$ and $R(\rho_2, W)$ values for those channels for fixed $\rho_1 \in [0, 1]$ and $\rho_2 \in [0, 1]$. Then, assuming $R(\rho_2, W) = R(\rho_2, W_{\mathsf{BEC}}) = R(\rho_2, W_{\mathsf{BSC}})$, we found the corresponding binary erasure channel and binary symmetric channel parameters. Finally, we checked whether any non-extremal (ρ_1, ρ_2) pairs for the binary erasure channel and the binary symmetric channel exist from the comparisons of the computed $E_0(\rho_1, W)$, $E_0(\rho_1, W_{\mathsf{BEC}})$, and $E_0(\rho_1, W_{\mathsf{BSC}})$.

Figure 4.1 displays the results of the experiment for minimum output alphabet size of 2 and maximum output alphabet size of 10. The non extremal (ρ_1, ρ_2) pairs are plotted in the ρ_1 versus ρ_2 plane.

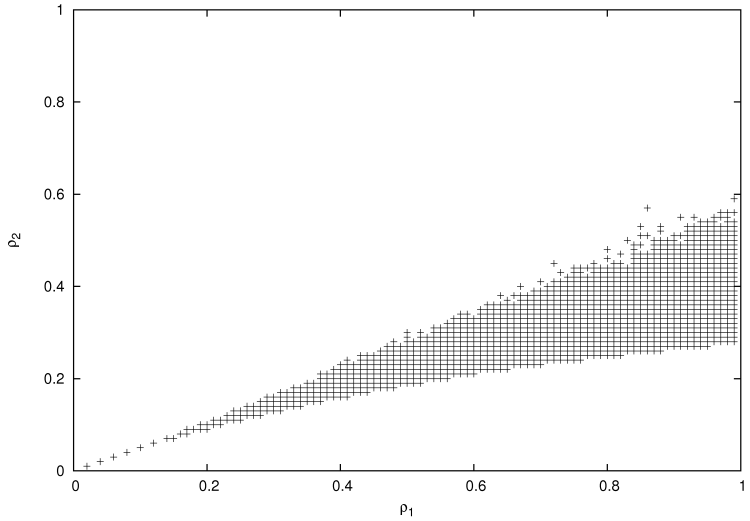

Figure 4.1: Numerical Experiment: Non extremal (ρ_1, ρ_2) pairs

At this point, we have to mention some possible drawbacks in the experiment.

Clearly, the output alphabet size is restricted. Results are valid only for the chosen maximum size of the output alphabet. Similarly, there might be some channels that we never generate and check. Finally, no analytic expressions exist to find the binary symmetric channel parameter. Hence, we use a version of the midpoint method that works within a specified precision error.

Despite some weaknesses, the numerical experiment results suggest that the extremal region in the ρ_1 versus ρ_2 plane can be extended beyond the case $\rho_1 = \rho_2 = \rho$.

Indeed, the theoretical result stated in the next theorem partially agree with the numerical experiment results. We give the proof in Appendix D.

Proposition 4.10. *Given a channel W, a binary symmetric channel W_{BSC}, and a binary erasure channel W_{BEC} such that*

$$E_0(\rho_1, W) = E_0(\rho_1, W_{\mathsf{BEC}}) = E_0(\rho_1, W_{\mathsf{BSC}})$$

holds for any fixed value of $\rho_1 \in [0, 1]$. Then, for $\rho_2 \in [0, 1]$, we have

$$R(\rho_2, W_{\mathsf{BSC}}) \leq R(\rho_2, W) \leq R(\rho_2, W_{\mathsf{BEC}}) \quad \textit{if} \quad \rho_2 \geq \rho_1 \tag{4.12}$$

Chapter 5

Conclusions

In this work, we studied channel polarization and polar codes, proposed by Arıkan in [1], as a recent subject in communication theory. Based on the properties we summarized in the introduction, two major characteristics of channel polarization and polar codes shaped this project.

First, the symmetric capacity of a communication channel is a fundamental quantity as a limit on the achievable rate of a communication system. On the other hand, we saw that the Bhattacharyya parameter is brought up as an auxiliary quantity to derive the main results on channel polarization and polar codes. Indeed, the proof of the fact that channel polarization arises from the recursive application of the basic channel transformations is easily carried by the combined analysis of the random processes corresponding to these two parameters. As a result, polar codes are shown to achieve the symmetric capacity of any B-DMC. Moreover, many properties about polar codes can be interchangeably stated using either one of the two parameters. Fortunately, these two parameters are not arbitrary. They turn out to be special cases of a more general channel parameter $E_0(\rho, W)$ we discussed in detail.

Second, we noted that the binary erasure channel and the binary symmetric channel represents extremal channels with respect to the Bhattacharyya parameters obtained after applying the basic channel transformations with a given initial value of the Bhattacharyya parameter. Furthermore, we observed that extremal polarization processes are used to derive results on both the rate of channel polarization [4], and the compound capacity of polar codes under successive cancellation decoding [6].

These observations lead us to study the function $E_0(\rho, W)$ and the random coding exponent $E_r(R)$, both defined by Gallager in [2], as more general channel parameters in the view of channel polarization. In particular, we showed that among all B-DMCs the binary erasure channel and the binary symmetric channel are extremal in the evolution of $E_0(\rho, W)$ under the basic channel transformations. In addition, we extended the extremality result in [3] using the parametric representation of the random coding exponent.

A

Appendix A

We derive some useful properties of the function $g(\rho, u)$ defined in (4.2) as

$$g(\rho, u) = \left(\frac{1}{2} (1+u)^{\frac{1}{1+\rho}} + \frac{1}{2} (1-u)^{\frac{1}{1+\rho}} \right)^{1+\rho}$$

Taking derivatives with respect to variable u, we have

$$\begin{aligned}
\frac{\partial g(\rho, u)}{\partial u} &= \frac{\partial}{\partial u} \quad \exp\{(1+\rho) \log \left(\frac{1}{2}(1+u)^{\frac{1}{1+\rho}} + \frac{1}{2}(1-u)^{\frac{1}{1+\rho}} \right)\} \\
&= \left(\frac{1}{2}(1+u)^{\frac{1}{1+\rho}} + \frac{1}{2}(1-u)^{\frac{1}{1+\rho}} \right)^{1+\rho} \left((1+\rho) \frac{\frac{1}{2}\frac{1}{1+\rho}(1+u)^{\frac{-\rho}{1+\rho}} - \frac{1}{2}\frac{1}{1+\rho}(1-u)^{\frac{-\rho}{1+\rho}}}{\frac{1}{2}(1+u)^{\frac{1}{1+\rho}} + \frac{1}{2}(1-u)^{\frac{1}{1+\rho}}} \right) \\
&= \left(\frac{1}{2} \right)^{1+\rho} \underbrace{\left(1 + \left(\frac{1-u}{1+u} \right)^{\frac{1}{1+\rho}} \right)^{\rho}}_{\geq 0} \underbrace{\left(1 - \left(\frac{1-u}{1+u} \right)^{\frac{-\rho}{1+\rho}} \right)}_{\leq 0} \leq 0
\end{aligned}$$

and

$$\frac{\partial^2 g(\rho, u)}{\partial u^2} = -\frac{\rho}{1+\rho} \left(1-u^2\right)^{\frac{1}{1+\rho}-2} \left(\frac{1}{2}(1+u)^{\frac{1}{1+\rho}} + \frac{1}{2}(1-u)^{\frac{1}{1+\rho}} \right)^{-1+\rho} \leq 0$$

Hence, $g(\rho, u)$ is a concave decreasing function in u. To simplify notation, we define

$$f(u) = \frac{1-u}{1+u} \tag{A.1}$$

$$\alpha(\rho, u) = (1 + f(u)^{\frac{1}{1+\rho}})^{\rho} \geq 0 \tag{A.2}$$

$$\beta(\rho, u) = (1 - f(u)^{\frac{-\rho}{1+\rho}}) \leq 0 \tag{A.3}$$

Then

$$\frac{\partial g(\rho, u)}{\partial u} = \frac{1}{2}^{1+\rho} \alpha(\rho, u) \beta(\rho, u)$$

Similarly, if we have a function $h(z,u)$ instead of u, we can get the following results

$$\begin{aligned}\frac{\partial g(\rho,h(z,u))}{\partial u} &= g'(\rho,h(z,u))h'(z,u)\\ &= \left(\frac{1}{2}\right)^{1+\rho}\alpha(\rho,h(z,u))\beta(\rho,h(z,u))h'(z,u)\end{aligned}$$

We next derive some expressions that we make use later in Appendix B.

$$\frac{\partial \alpha(\rho,h(z,u))}{\partial u} = \frac{\rho}{1+\rho}h'(z,u)f'(h(z,u))f(h(z,u))^{\frac{-\rho}{1+\rho}}(1+f(h(z,u))^{\frac{1}{1+\rho}})^{\rho-1} \tag{A.4}$$

$$\frac{\partial \beta(\rho,h(z,u))}{\partial u} = \frac{\rho}{1+\rho}h'(z,u)f'(h(z,u))f(h(z,u))^{\frac{-\rho}{1+\rho}-1} \tag{A.5}$$

where

$$f'(u) = \frac{-2}{(1+u)^2}.$$

We define the function

$$F_{z,\rho_1,\rho_2}(t) = g(\rho_2, h(z, g^{-1}(\rho_1,t))). \tag{A.6}$$

Assume we are interested with the convexity of $F_{z,\rho_1,\rho_2}(t)$ with respect to variable t. Taking the first derivative we obtain

$$\begin{aligned}\frac{\partial F_{z,\rho_1,\rho_2}(t)}{\partial t} &= \frac{\partial}{\partial t}g(\rho_2,h(z,g^{-1}(\rho_1,t)))\\ &= \frac{g'(\rho_2,h(z,g^{-1}(\rho_1,t)))}{g'(\rho_1,g^{-1}(\rho_1,t))}h'(z,g^{-1}(\rho_1,t)).\end{aligned}$$

We define $u = g^{-1}(\rho_1,t)$. Since $g(\rho,u)$ is a decreasing function in u, so is $g^{-1}(\rho_1,t)$ in t. Hence we can check the convexity of $F_{z,\rho_1,\rho_2}(t)$ with respect to the variable t, from the monotonicity with respect to u of the following expression:

$$h'(z,u)\frac{g'(\rho_2,h(z,u))}{g'(\rho_1,u)} = 2^{\rho_1-\rho_2}h'(z,u)\frac{\alpha(\rho_2,h(z,u))\beta(\rho_2,h(z,u))}{\alpha(\rho_1,u)\beta(\rho_1,u)}. \tag{A.7}$$

B

Appendix B

In this appendix, we show that the function $F_{z,\rho}(t)$ defined in (4.6) is convex with respect to variable t for fixed $\rho \in [0,1]$ and $z \in [0,1]$ values. From Appendix A, we know that we are interested in the convexity of $F_{z,\rho_1,\rho_2}(t) = g(\rho_2, h(z, g^{-1}(\rho_1, t)))$ defined in (A.6) in the particular case when $\rho_1 = \rho_2 = \rho$ and $h(z,u) = zu$. By (A.7), we need to check the monotonicity of the following expression in u

$$\begin{aligned} h'(z,u)\frac{g'(\rho_2, h(z,u))}{g'(\rho_1,u)} &= 2^{\rho_1-\rho_2} h'(z,u)\frac{\alpha(\rho_2,h(z,u))\beta(\rho_2,h(z,u))}{\alpha(\rho_1,u)\beta(\rho_1,u)} \\ &= z\frac{\alpha(\rho,h(z,u))\beta(\rho,h(z,u))}{\alpha(\rho,u)\beta(\rho,u)} \end{aligned} \tag{B.1}$$

where the functions $\alpha(\rho,u)$ and $\beta(\rho,u)$ are defined in (A.2) and (A.3), respectively. Now taking the derivative of (B.1) with respect to u, we get

$$\begin{aligned} &\frac{\partial}{\partial u} \quad z\frac{\alpha(\rho,h(z,u))\beta(\rho,h(z,u))}{\alpha(\rho,u)\beta(\rho,u)} \\ =&z \underbrace{\frac{\alpha(\rho,h(z,u))\beta(\rho,h(z,u))}{\alpha(\rho,u)\beta(\rho,u)}}_{\geq 0} \\ &\left(\frac{\partial\alpha(\rho,h(z,u))/\partial u}{\alpha(\rho,h(z,u))} + \frac{\partial\beta(\rho,h(z,u))/\partial u}{\beta(\rho,h(z,u))} - \frac{\partial\alpha(\rho,u)/\partial u}{\alpha(\rho,u)} - \frac{\partial\beta(\rho,u)/\partial u}{\beta(\rho,u)}\right) \end{aligned} \tag{B.2}$$

We can see that the sign of the expression inside the parenthesis in (B.2) will determine whether the expression in (B.1) is increasing or decreasing in u. At this point, we note that

$$\frac{\partial\alpha(\rho,u)/\partial u}{\alpha(\rho,u)} + \frac{\partial\beta(\rho,u)/\partial u}{\beta(\rho,u)} = \left(\frac{\partial\alpha(\rho,h(z,u))/\partial u}{\alpha(\rho,h(z,u))} + \frac{\partial\beta(\rho,h(z,u))/\partial u}{\beta(\rho,h(z,u))}\right)\Big|_{z=1} \tag{B.3}$$

Moreover, we claim that the expression inside the paranthesis in the RHS of (B.3) is increasing in z. As a consequence, $F_{z,\rho}(t)$ is a concave function in $u = g^{-1}(\rho,t)$. Since u is decreasing in t, we have

$$\frac{\partial F'_{z,\rho}(u)}{\partial u}\frac{\partial u}{\partial t} \geq 0.$$

We conclude that $F_{z,\rho}(t)$ is a convex function with respect to variable t.

In the rest of the appendix, we prove our claim. Using the equations (A.4) and (A.5) from Appendix A, we have

$$
\begin{aligned}
&\frac{\partial\alpha(\rho,h(z,u))/\partial u}{\alpha(\rho,h(z,u))}+\frac{\partial\beta(\rho,h(z,u))/\partial u}{\beta(\rho,h(z,u))}\\
=\;&\frac{\rho}{1+\rho}f(zu)^{\frac{-\rho}{1+\rho}-1}zf'(zu)\left(\frac{f(zu)}{1+f(zu)^{\frac{1}{1+\rho}}}+\frac{1}{1-f(zu)^{\frac{-\rho}{1+\rho}}}\right)\\
=\;&\frac{\rho}{1+\rho}f(zu)^{\frac{-\rho}{1+\rho}-1}zf'(zu)\left(\frac{f(zu)-f(zu)^{\frac{1}{1+\rho}}+1+f(zu)^{\frac{1}{1+\rho}}}{(1+f(zu)^{\frac{1}{1+\rho}})(1-f(zu)^{\frac{-\rho}{1+\rho}})}\right)\\
=\;&\frac{\rho}{1+\rho}f(zu)^{\frac{-\rho}{1+\rho}-1}zf'(zu)(1+f(zu))(1+f(zu)^{\frac{1}{1+\rho}})^{-1}(1-f(zu)^{\frac{-\rho}{1+\rho}})^{-1}\\
=\;&\frac{\rho}{1+\rho}zf'(zu)(1+f(zu)^{-1})(1+f(zu)^{\frac{1}{1+\rho}})^{-1}(f(zu)^{\frac{\rho}{1+\rho}}-1)^{-1}\\
=\;&\frac{\rho}{1+\rho}\frac{-4z}{(1+zu)^2(1-zu)}\left(1+\left(\frac{1-zu}{1+zu}\right)^{\frac{1}{1+\rho}}\right)^{-1}\left(-1+\left(\frac{1-zu}{1+zu}\right)^{\frac{\rho}{1+\rho}}\right)^{-1}\\
=\;&\frac{4\rho}{1+\rho}\left(\underbrace{\frac{1-z^2u^2}{z}\left((1+zu)^{\frac{\rho}{1+\rho}}-(1-zu)^{\frac{\rho}{1+\rho}}\right)}_{\text{Part 2}}\underbrace{\left((1+zu)^{\frac{1}{1+\rho}}+(1-zu)^{\frac{1}{1+\rho}}\right)}_{\text{Part 1}}\right)^{-1}
\end{aligned}
$$

We consider the expressions labeled as Part 1 and Part 2 separately. Note that both are positive valued. In addition, as we next show, both are decreasing in z. As a result, we deduce

$$
\begin{aligned}
&\frac{\partial}{\partial z}\left(\frac{(1-z^2u^2)}{z}\left((1+zu)^{\frac{1}{1+\rho}}+(1-zu)^{\frac{1}{1+\rho}}\right)\left((1+uz)^{\frac{\rho}{1+\rho}}-(1-uz)^{\frac{\rho}{1+\rho}}\right)\right)\leq 0\\
&\frac{\partial}{\partial z}\left(\frac{(1-z^2u^2)}{z}\left((1+zu)^{\frac{1}{1+\rho}}+(1-zu)^{\frac{1}{1+\rho}}\right)\left((1+uz)^{\frac{\rho}{1+\rho}}-(1-uz)^{\frac{\rho}{1+\rho}}\right)\right)^{-1}\geq 0
\end{aligned}
$$

which is indeed the claim we want to prove.

For Part 1, we get

$$
\begin{aligned}
&\frac{\partial}{\partial z}\left((1+zu)^{\frac{1}{1+\rho}}+(1-zu)^{\frac{1}{1+\rho}}\right)\\
=&\frac{u\left((1+uz)^{\frac{-\rho}{1+\rho}}-(1-uz)^{\frac{-\rho}{1+\rho}}\right)}{1+\rho}\leq 0
\end{aligned}
$$

For Part 2, we have

$$
\begin{aligned}
&\frac{\partial}{\partial z}\left(\frac{(1-u^2z^2)}{z}\left((1+uz)^{\frac{\rho}{1+\rho}}-(1-uz)^{\frac{\rho}{1+\rho}}\right)\right)\\
=&\ \frac{1}{z^2}\frac{\rho uz\,(1-u^2z^2)\left((1+uz)^{\frac{\rho}{1+\rho}-1}+(1-uz)^{\frac{\rho}{1+\rho}-1}\right)}{1+\rho}\\
&+\frac{1}{z^2}\left(1+u^2z^2\right)\left(-(1+uz)^{\frac{\rho}{1+\rho}}+(1-uz)^{\frac{\rho}{1+\rho}}\right)\\
=&\frac{1}{z^2}\Bigg(\ (1+uz)^{\frac{\rho}{1+\rho}}\left(\frac{\rho}{1+\rho}uz\,(1-uz)-(1+u^2z^2)\right)\\
&\qquad+(1-uz)^{\frac{\rho}{1+\rho}}\left(\frac{\rho}{1+\rho}uz\,(1+uz)+(1+u^2z^2)\right)\Bigg)\\
=&\frac{1}{z^2}\Big(\ -(1+x)^k\left((k+1)\,x^2-kx+1\right)+(1-x)^k\left((k+1)\,x^2+kx+1\right)\Big)\\
=&\frac{1}{z^2}\left(\ -f_1(x,k)+f_2(x,k)\right) \qquad \text{(B.4)}
\end{aligned}
$$

where $k=\frac{\rho}{1+\rho}\in(0,\frac{1}{2})$, $x=uz\in(0,1)$, and

$$f_1(x,k)=(1+x)^k\left((k+1)\,x^2-kx+1\right) \qquad \text{(B.5)}$$

$$f_2(x,k)=(1-x)^k\left((k+1)\,x^2+kx+1\right). \qquad \text{(B.6)}$$

We immediately observe that $f_1(x,k)=f_2(x,k)$ when $k=0$. We now show that $-f_1(x,k)+f_2(x,k)\le 0$ for all x and k. Taking the derivatives of $f_1(x,k)$ and $f_2(x,k)$ with respect to k, we obtain

$$
\begin{aligned}
&\frac{\partial f_1(x,k)}{\partial k}=\frac{\partial}{\partial k}(1+x)^k\left((k+1)\,x^2-kx+1\right)\\
=&\,(1+x)^k\log(1+x)\left((k+1)\,x^2-kx+1\right)+(1+x)^k\left(x^2-x\right)\\
=&\,(1+x)^k\left(\log(1+x)\left((k+1)\,x^2-kx+1\right)+x^2-x\right)\\
=&\,(1+x)^k\left(\sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n}\left((k+1)\,x^2-kx+1\right)+x^2-x\right)\\
=&\,(1+x)^k\left(\sum_{m=3}^{\infty}(-1)^{m+1}x^m\left(\frac{1+k}{m-2}+\frac{k}{m-1}+\frac{1}{m}\right)+x^2\left(\frac{1}{2}-k\right)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\partial f_2(x,k)}{\partial k}=\frac{\partial}{\partial k}(1-x)^k\left((k+1)\,x^2+kx+1\right)\\
=&\,(1-x)^k\log(1-x)\left((k+1)\,x^2+kx+1\right)+(1-x)^k\left(x^2+x\right)\\
=&\,(1-x)^k\left(-\sum_{n=1}^{\infty}\frac{x^n}{n}\left((k+1)\,x^2+kx+1\right)+x^2+x\right)\\
=&\,(1-x)^k\left(-\sum_{m=3}^{\infty}x^m\left(\frac{1+k}{m-2}+\frac{k}{m-1}+\frac{1}{m}\right)+x^2\left(\frac{1}{2}-k\right)\right)
\end{aligned}
$$

where we used the Taylor series expansions of $\log(1+x)$ and $\log(1-x)$. Moreover, we have

$$(1+x)^k \geq (1-x)^k \geq 0$$
$$\sum_{m=3}^{\infty} (-1)^{m+1} x^m \left(\frac{1+k}{m-2} + \frac{k}{m-1} + \frac{1}{m}\right) \geq -\sum_{m=3}^{\infty} x^m \left(\frac{1+k}{m-2} + \frac{k}{m-1} + \frac{1}{m}\right)$$
$$x^2 \left(\frac{1}{2} - k\right) \geq 0.$$

These imply that

$$\frac{\partial f_1(x,k)}{\partial k} \geq \frac{\partial f_2(x,k)}{\partial k}. \tag{B.7}$$

Hence, the relation in (B.7) together with the fact that $f_1(x,k) = f_2(x,k)$ when $k=0$, proves that $-f_1(x,k) + f_2(x,k) \leq 0$ for all x and k. Consequently, Part 2 is also decreasing in z

$$\frac{\partial}{\partial z}\left(\frac{(1-u^2z^2)}{z}\left((1+uz)^{\frac{\rho}{1+\rho}} - (1-uz)^{\frac{\rho}{1+\rho}}\right)\right) \leq 0.$$

This proves our claim that the RHS of (B.3) is increasing in z.

C

Appendix C

We show that the function $t_3(z_1, z_2)$ given in (4.8) is a decreasing concave function in both z_1 and z_2 separately. For simplicity, we first define

$$\begin{aligned} g(\rho, u) &= g(u) \\ t_1(z_1, z_2) &= (1 + z_1 z_2)\ g(\frac{z_1 + z_2}{1 + z_1 z_2}) \\ t_2(z_1, z_2) &= (1 - z_1 z_2)\ g(\frac{z_1 - z_2}{1 - z_1 z_2}) \end{aligned}$$

Hence, we get

$$t_3(z_1, z_2) = \frac{1}{2} t_1(z_1, z_2) + \frac{1}{2} t_2(z_1, z_2)$$

Let g'_z and g''_z denote the first and second derivatives with respect to variable z. Then, taking derivatives of $t_1(z_1, z_2)$ with respect to z_1, we get

$$\begin{aligned} \frac{\partial}{\partial z_1} t_1(z_1, z_2) &= z_2\ g(\frac{z_1 + z_2}{1 + z_1 z_2}) + \frac{1 - {z_2}^2}{1 + z_1 z_2}\ g'_{z_1}(\frac{z_1 + z_2}{1 + z_1 z_2}) \\ \frac{\partial^2}{\partial {z_1}^2} t_1(z_1, z_2) &= z_2\ g'_{z_1}(\frac{z_1 + z_2}{1 + z_1 z_2}) \frac{1 - {z_2}^2}{(1 + z_1 z_2)^2} - z_2 \frac{1 - {z_2}^2}{(1 + z_1 z_2)^2}\ g'_{z_1}(\frac{z_1 + z_2}{1 + z_1 z_2}) \\ &\quad + \frac{1 - {z_2}^2}{1 + z_1 z_2}\ g''_{z_1}(\frac{z_1 + z_2}{1 + z_1 z_2}) \frac{1 - {z_2}^2}{(1 + z_1 z_2)^2} \\ &= \frac{(1 - {z_2}^2)^2}{(1 + z_1 z_2)^3}\ g''_{z_1}(\frac{z_1 + z_2}{1 + z_1 z_2}) \end{aligned}$$

By symmetry, we deduce the following derivatives of $t_1(z_1, z_2)$ with respect to z_2

$$\begin{aligned} \frac{\partial}{\partial z_2} t_1(z_1, z_2) &= z_1\ g(\frac{z_1 + z_2}{1 + z_1 z_2}) + \frac{1 - {z_1}^2}{1 + z_1 z_2}\ g'_{z_2}(\frac{z_1 + z_2}{1 + z_1 z_2}) \\ \frac{\partial^2}{\partial {z_1}^2} t_1(z_1, z_2) &= \frac{(1 - {z_1}^2)^2}{(1 + z_1 z_2)^3}\ g''_{z_2}(\frac{z_1 + z_2}{1 + z_1 z_2}) \end{aligned}$$

Similarly, we can obtain the derivatives of $t_2(z_1, z_2)$ with respect to z_1 as

$$\begin{aligned} \frac{\partial}{\partial z_1} t_2(z_1, z_2) &= -z_2\ g(\frac{z_1 - z_2}{1 - z_1 z_2}) + \frac{1 - {z_2}^2}{1 - z_1 z_2}\ g'_{z_1}(\frac{z_1 - z_2}{1 - z_1 z_2}) \\ \frac{\partial^2}{\partial {z_1}^2} t_2(z_1, z_2) &= \frac{(1 - {z_2}^2)^2}{(1 - z_1 z_2)^3}\ g''_{z_1}(\frac{z_1 - z_2}{1 - z_1 z_2}) \end{aligned}$$

and with respect to z_2 as

$$\frac{\partial}{\partial z_2} t_2(z_1, z_2) = -z_1 \;\; g(\frac{z_1 - z_2}{1 - z_1 z_2}) - \frac{1 - {z_1}^2}{1 - z_1 z_2} \;\; g'_{z_2}(\frac{z_1 - z_2}{1 - z_1 z_2})$$
$$\frac{\partial^2}{\partial {z_1}^2} t_1(z_1, z_2) = \frac{(1 - {z_1}^2)^2}{(1 - z_1 z_2)^3} \;\; g''_{z_2}(\frac{z_1 - z_2}{1 - z_1 z_2})$$

So, we get

$$\begin{aligned}\frac{\partial}{\partial z_1} t_3(z_1, z_2) = &\frac{1}{2} z_2 \;\; g(\frac{z_1 + z_2}{1 + z_1 z_2}) + \frac{1 - {z_2}^2}{2(1 + z_1 z_2)} \;\; g'_{z_1}(\frac{z_1 + z_2}{1 + z_1 z_2}) \\ &- \frac{1}{2} z_2 \;\; g(\frac{z_1 - z_2}{1 - z_1 z_2}) + \frac{1 - {z_2}^2}{2(1 - z_1 z_2)} \;\; g'_{z_1}(\frac{z_1 - z_2}{1 - z_1 z_2}) \leq 0 \\ \frac{\partial}{\partial z_2} t_3(z_1, z_2) = &\frac{1}{2} z_1 \;\; g(\frac{z_1 + z_2}{1 + z_1 z_2}) + \frac{1 - {z_1}^2}{2(1 + z_1 z_2)} \;\; g'_{z_2}(\frac{z_1 + z_2}{1 + z_1 z_2}) \\ &- \frac{1}{2} z_1 \;\; g(\frac{z_1 - z_2}{1 - z_1 z_2}) + \frac{1 - {z_1}^2}{2(1 - z_1 z_2)} \;\; g'_{z_2}(\frac{z_1 - z_2}{1 - z_1 z_2})\end{aligned}$$

where the negativity follows since the function $g(\rho, u)$ is positive, decreasing in u by Appendix A, and for any fixed $z_1, z_2 \in [0, 1]$

$$\frac{z_1 + z_2}{1 + z_1 z_2} \geq \frac{z_1 - z_2}{1 - z_1 z_2} \Rightarrow g(\frac{z_1 + z_2}{1 + z_1 z_2}) \leq g(\frac{z_1 - z_2}{1 - z_1 z_2})$$

holds. As a result, $t_3(z_1, z_2)$ is decreasing in both z_1, and z_2 separetely.
On the other hand, the concavity of $t_3(z_1, z_2)$ can be easily deduced from

$$\frac{\partial^2}{\partial {z_1}^2} t_3(z_1, z_2) = \frac{1}{2} \frac{(1 - {z_2}^2)^2}{(1 + z_1 z_2)^3} \;\; g''_{z_1}(\frac{z_1 + z_2}{1 + z_1 z_2}) + \frac{1}{2} \frac{(1 - {z_2}^2)^2}{(1 - z_1 z_2)^3} \;\; g''_{z_1}(\frac{z_1 - z_2}{1 - z_1 z_2}) \leq 0$$
$$\frac{\partial^2}{\partial {z_2}^2} t_3(z_1, z_2) = \frac{1}{2} \frac{(1 - {z_1}^2)^2}{(1 + z_1 z_2)^3} \;\; g''_{z_2}(\frac{z_1 + z_2}{1 + z_1 z_2}) + \frac{1}{2} \frac{(1 - {z_1}^2)^2}{(1 - z_1 z_2)^3} \;\; g''_{z_2}(\frac{z_1 - z_2}{1 - z_1 z_2}) \leq 0$$

as we already showed in Appendix A that the function $g(u)$ is a concave function with respect to variable $u \in [0, 1]$.

D

Appendix D

Given a channel W with $E_0(\rho_1, W)$ for $0 \leq \rho_1 \leq 1$, we want to know when the binary erasure channel W_{BEC} and the binary symmetric channel W_{BSC} defined through the equality

$$E_0(\rho_1, W) = E_0(\rho_1, W_{\mathsf{BEC}}) = E_0(\rho_1, W_{\mathsf{BSC}})$$

are extremal for the value of $R(\rho_2, W)$ where $0 \leq \rho_2 \leq 1$. By the definition in (2.4) and Lemma 4.1, we have

$$R(\rho_2, W) = \frac{\partial E_0(\rho_2, W)}{\partial \rho_2} = \frac{\mathbb{E}[-\partial g(\rho_2, Z)/\partial \rho_2]}{\mathbb{E}[g(\rho_2, Z)]}$$

We want to check convexity of the function $\partial g(\rho_2, g^{-1}(\rho_1, t))/\partial \rho_2$ with respect to variable t for fixed values of ρ_1 and ρ_2. This corresponds to checking the convexity of the function

$$\frac{\partial F_{1,\rho_1,\rho_2}(t)}{\partial \rho_2} = \frac{\partial}{\partial \rho_2} g(\rho_2, h(z, g^{-1}(\rho_1, t))) \tag{D.1}$$

where $h(z, u) = u$ for fixed values of ρ_1 and ρ_2. From Appendix A, we know we can equivalently check the monotonicity of the following expression

$$\begin{aligned}
&\frac{\partial}{\partial \rho_2} h'(1,u) \frac{g'(\rho_2, h(1,u))}{g'(\rho_1, u)} \\
=&\frac{\partial}{\partial \rho_2} 2^{\rho_1 - \rho_2} h'(1,u) \frac{\alpha(\rho_2, h(1,u))\beta(\rho_2, h(1,u))}{\alpha(\rho_1, u)\beta(\rho_1, u)} \\
=&\frac{\partial}{\partial \rho_2} \frac{2^{-\rho_2}\alpha(\rho_2, h(1,u))\beta(\rho_2, h(1,u))}{2^{-\rho_1}\alpha(\rho_1, u)\beta(\rho_1, u)} \\
=&\frac{2^{-\rho_2}\alpha(\rho_2, h(1,u))\beta(\rho_2, h(1,u))}{2^{-\rho_1}\alpha(\rho_1, u)\beta(\rho_1, u)} \left(\frac{\partial 2^{-\rho_2}\alpha(\rho_2, h(1,u))/\partial \rho 2}{2^{-\rho_2}\alpha(\rho_2, h(1,u))} + \frac{\partial \beta(\rho_2, h(1,u))/\partial \rho_2}{\beta(\rho_2, h(1,u))} \right) \\
=&\Phi(u, \rho_1, \rho_2)\Psi(u, \rho_2)
\end{aligned} \tag{D.2}$$

where

$$\begin{aligned}
\Phi(u, \rho_1, \rho_2) &= \frac{2^{-\rho_2}\alpha(\rho_2, h(1,u))\beta(\rho_2, h(1,u))}{2^{-\rho_1}\alpha(\rho_1, u)\beta(\rho_1, u)} \geq 0 \\
\Psi(u, \rho_2) &= \left(\frac{\partial 2^{-\rho_2}\alpha(\rho_2, h(1,u))/\partial \rho 2}{2^{-\rho_2}\alpha(\rho_2, h(1,u))} + \frac{\partial \beta(\rho_2, h(1,u))/\partial \rho 2}{\beta(\rho_2, h(1,u))} \right)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial}{\partial \rho_2} 2^{-\rho_2}\alpha(\rho_2, h(1,u)) &= \frac{\partial}{\partial \rho_2}\left(\frac{1}{2}+\frac{1}{2}f(u)^{\frac{1}{1+\rho_2}}\right)^{\rho_2} \\
&= \left(\frac{1}{2}+\frac{1}{2}f(u)^{\frac{1}{1+\rho_2}}\right)^{\rho_2} \\
&\quad \left(\log\left(\frac{1}{2}+\frac{1}{2}f(u)^{\frac{1}{1+\rho_2}}\right) + \rho_2 \frac{\frac{1}{2}f(t)^{\frac{1}{1+\rho_2}}\frac{-1}{(1+\rho_2)^2}\log f(t)}{\frac{1}{2}+\frac{1}{2}f(u)^{\frac{1}{1+\rho_2}}}\right) \\
&= 2^{-\rho_2}\alpha(\rho_2, h(1,u)) \\
&\quad \left(\log\left(\frac{1}{2}+\frac{1}{2}f(u)^{\frac{1}{1+\rho_2}}\right) - \frac{\rho_2 f(t)^{\frac{1}{1+\rho_2}}}{(1+\rho_2)\left(1+f(t)^{\frac{1}{1+\rho_2}}\right)}\log f(t)\right) \\
\frac{\partial}{\rho_2}\beta(\rho_2, h(1,u)) &= \frac{\partial}{\rho_2}\left(1-f(u)^{\frac{-\rho_2}{1+\rho_2}}\right) \\
&= \frac{1}{(1+\rho_2)^2} f(u)^{\frac{-\rho_2}{1+\rho_2}}\log f(u)
\end{aligned}
$$

Hence

$$
\Phi(u,\rho_1,\rho_2) = \frac{\left(\frac{1}{2}+\frac{1}{2}f(u)^{\frac{1}{1+\rho_2}}\right)^{\rho_2}\left(1-f(u)^{\frac{-\rho_2}{1+\rho_2}}\right)}{\left(\frac{1}{2}+\frac{1}{2}f(u)^{\frac{1}{1+\rho_1}}\right)^{\rho_1}\left(1-f(u)^{\frac{-\rho_1}{1+\rho_1}}\right)} \tag{D.3}
$$

and

$$
\begin{aligned}
\Psi(u,\rho_2) = &\log\left(\frac{1}{2}+\frac{1}{2}f(u)^{\frac{1}{1+\rho_2}}\right) \\
&+ \frac{1}{(1+\rho_2)^2}\log f(u)\left(-\rho_2\frac{f(u)^{\frac{1}{1+\rho_2}}}{1+f(u)^{\frac{1}{1+\rho_2}}} + \frac{1}{f(u)^{\frac{\rho_2}{1+\rho_2}}-1}\right)
\end{aligned} \tag{D.4}
$$

To simplify derivations we define $k = f(u)$. Taking derivative with respect to k

$$
\begin{aligned}
\frac{\partial}{\partial k}\Phi(k,\rho_1,\rho_2)\Psi(k,\rho_2) &= \Phi'(k,\rho_1,\rho_2)\Psi(k,\rho_2) + \Phi(k,\rho_1,\rho_2)\Psi'(k,\rho_2) \\
&= \Phi(k,\rho_1,\rho_2)\Psi(k,\rho_2)\left(\frac{\partial \log\Phi(k,\rho_1,\rho_2)}{\partial k} + \frac{\Psi'(k,\rho_2)}{\Psi(k,\rho_2)}\right)
\end{aligned}
$$

where we abuse notation to redefine functions as

$$
\begin{aligned}
\Phi(k,\rho_1,\rho_2) &= \frac{\left(\frac{1}{2}+\frac{1}{2}k^{\frac{1}{1+\rho_2}}\right)^{\rho_2}\left(1-k^{\frac{-\rho_2}{1+\rho_2}}\right)}{\left(\frac{1}{2}+\frac{1}{2}k^{\frac{1}{1+\rho_1}}\right)^{\rho_1}\left(1-k^{\frac{-\rho_1}{1+\rho_1}}\right)} \\
\Psi(k,\rho_2) &= \log\left(\frac{1}{2}+\frac{1}{2}k^{\frac{1}{1+\rho_2}}\right) + \frac{1}{(1+\rho_2)^2}\left(\frac{1+k^{\frac{1}{1+\rho_2}}-\rho_2\left(k-k^{\frac{1}{1+\rho_2}}\right)}{\left(1+k^{\frac{1}{1+\rho_2}}\right)\left(k^{\frac{\rho_2}{1+\rho_2}}-1\right)}\right)\log k \\
&= \log\left(\frac{1}{2}+\frac{1}{2}k^{\frac{1}{1+\rho_2}}\right) + \frac{\left(1+k^{\frac{1}{1+\rho_2}}-\rho_2\left(k-k^{\frac{1}{1+\rho_2}}\right)\right)\log k}{(1+\rho_2)^2 \quad \gamma(k,\rho_2)}
\end{aligned}
$$

where

$$\gamma(k,\rho_2) = \left(1 + k^{\frac{1}{1+\rho_2}}\right)\left(k^{\frac{\rho_2}{1+\rho_2}} - 1\right) \leq 0 \tag{D.5}$$

We now derive the expressions in the above equation

$$\begin{aligned}
\log \Phi(k,\rho_1,\rho_2) = \;& \rho_2 \log\left(\frac{1}{2} + \frac{1}{2}k^{\frac{1}{1+\rho_2}}\right) + \log\left(1 - k^{\frac{-\rho_2}{1+\rho_2}}\right) \\
& - \rho_1 \log\left(\frac{1}{2} + \frac{1}{2}k^{\frac{1}{1+\rho_1}}\right) - \log\left(1 - k^{\frac{-\rho_1}{1+\rho_1}}\right) \\
\Rightarrow \frac{\partial \log \Phi(k,\rho_1,\rho_2)}{\partial k} = & \frac{\rho_2}{1+\rho_2}\frac{k^{\frac{-\rho_2}{1+\rho_2}}}{1 + k^{\frac{1}{1+\rho_2}}} + \frac{\rho_2}{1+\rho_2}\frac{k^{\frac{-\rho_2}{1+\rho_2}-1}}{1 - k^{\frac{-\rho_2}{1+\rho_2}}} \\
& - \frac{\rho_1}{1+\rho_1}\frac{k^{\frac{-\rho_1}{1+\rho_1}}}{1 + k^{\frac{1}{1+\rho_1}}} - \frac{\rho_1}{1+\rho_1}\frac{k^{\frac{-\rho_1}{1+\rho_1}-1}}{1 - k^{\frac{-\rho_1}{1+\rho_1}}} \\
= & \frac{\rho_2}{1+\rho_2}\frac{1+k}{k\left(1 + k^{\frac{1}{1+\rho_2}}\right)\left(k^{\frac{\rho_2}{1+\rho_2}} - 1\right)} - \frac{\rho_1}{1+\rho_1}\frac{1+k}{k\left(1 + k^{\frac{1}{1+\rho_1}}\right)\left(k^{\frac{\rho_1}{1+\rho_1}} - 1\right)} \\
= & F(k,\rho_2) - F(k,\rho_1)
\end{aligned}$$

where

$$F(k,\rho) = \frac{\rho}{1+\rho}\frac{1+k}{k}\frac{1}{\gamma(k,\rho)} \tag{D.6}$$

and

$$\begin{aligned}
\Psi'(k,\rho_2) = \;& \frac{1}{1+\rho_2}\frac{k^{-\frac{\rho_2}{1+\rho_2}}}{1 + k^{\frac{1}{1+\rho_2}}} + \frac{1}{(1+\rho_2)^2}\frac{1}{k}\left(-\rho_2\frac{k^{\frac{1}{1+\rho_2}}}{1 + k^{\frac{1}{1+\rho_2}}} + \frac{1}{k^{\frac{\rho_2}{1+\rho_2}} - 1}\right) \\
& + \frac{1}{(1+\rho_2)^2}\log k\left(\frac{-\rho_2}{1+\rho_2}\frac{k^{-\frac{\rho_2}{1+\rho_2}}}{\left(1 + k^{\frac{1}{1+\rho_2}}\right)^2} - \frac{\rho_2}{1+\rho_2}\frac{k^{\frac{-1}{1+\rho_2}}}{\left(k^{\frac{\rho_2}{1+\rho_2}} - 1\right)^2}\right) \\
= \;& \frac{1 - k^{\frac{-\rho_2}{1+\rho_2}}}{(1+\rho_2)\,\gamma(k,\rho_2)} + \frac{\left(1 + k^{\frac{1}{1+\rho_2}} - \rho_2\left(k - k^{\frac{1}{1+\rho_2}}\right)\right)}{(1+\rho_2)^2 \quad k \quad \gamma(k,\rho_2)} \\
& - \frac{\rho_2\,(k+1)\left(k^{\frac{\rho_2}{1+\rho_2}} + k^{\frac{1}{1+\rho_2}}\right)\log k}{(1+\rho_2)^3 \quad k \quad \gamma^2(k,\rho_2)} \\
= \;& \frac{k+1}{(1+\rho_2)^2 \;\; k \;\; \gamma(k,\rho_2)} - \frac{\rho_2\,(k+1)\left(k^{\frac{\rho_2}{1+\rho_2}} + k^{\frac{1}{1+\rho_2}}\right)\log k}{(1+\rho_2)^3 \;\; k \;\; \gamma^2(k,\rho_2)} \\
= \;& \frac{1}{\rho_2(1+\rho_2)}F(k,\rho_2)\left(1 - \frac{\rho_2\left(k^{\frac{\rho_2}{1+\rho_2}} + k^{\frac{1}{1+\rho_2}}\right)\log k}{(1+\rho_2)\left(1 + k^{\frac{1}{1+\rho_2}}\right)\left(k^{\frac{\rho_2}{1+\rho_2}} - 1\right)}\right)
\end{aligned} \tag{D.7}$$

Now, we show that $\Psi'(k,\rho_2) \leq 0$. Since $F(k,\rho_2) \leq 0$, it is sufficient to show that the term inside the paranthesis is positive, i.e.,

$$\frac{\rho_2 \left(k^{\frac{\rho_2}{1+\rho_2}} + k^{\frac{1}{1+\rho_2}}\right) \log k}{(1+\rho_2)\left(1+k^{\frac{1}{1+\rho_2}}\right)\left(k^{\frac{\rho_2}{1+\rho_2}}-1\right)} \leq 1. \tag{D.8}$$

This proof is carried in Appendix E.

Recall that we are interested in the sign of the following expression

$$\frac{\partial}{\partial k}\Phi(k,\rho_1,\rho_2)\Psi(k,\rho_2) = \Phi(k,\rho_1,\rho_2)\Psi(k,\rho_2)\left(\frac{\Psi'(k,\rho_2)}{\Psi(k,\rho_2)} + F(k,\rho_2) - F(k,\rho_1)\right). \tag{D.9}$$

We also proved $\Psi'(k,\rho) \leq 0$ holds. This implies that

$$\Psi(k,\rho) \geq \lim_{k\to 1}\Psi(1,\rho) = \frac{2}{(1+\rho)^2}\lim_{k\to 1}\frac{\log k}{\gamma(k,\rho)} = \frac{1}{\rho\,(1+\rho)} \geq 0$$

since

$$\lim_{k\to 1}\frac{\log k}{\gamma(k,\rho)} = \frac{0}{0} = \lim_{k\to 1}\frac{\partial \log k/\partial k}{\partial \gamma(k,\rho)/\partial k} = \lim_{k\to 1}\frac{k+\rho k}{k\left(k+\rho k - k^{\frac{1}{1+\rho}} + g k^{\frac{\rho}{1+\rho}}\right)} = \frac{1+\rho}{2\rho}.$$

We conclude that the function defined in (D.1) is concave with respect to t when $\rho_1 = \rho_2$ since

$$\underbrace{\Phi(k,\rho_1,\rho_2)}_{\geq 0}\Psi(k,\rho_2)\left(\frac{\Psi'(k,\rho_2)}{\Psi(k,\rho_2)} + F(k,\rho_2) - F(k,\rho_1)_{\rho_1=\rho_2}\right) \leq 0$$

holds. Therefore, we have

$$\frac{\Psi'(k,\rho_2)}{\Psi(k,\rho_2)} + F(k,\rho_2) \leq F(k,\rho_1)_{\rho_1=\rho_2}. \tag{D.10}$$

We observe that if the function $F(k,\rho)$ is monotonic in ρ, we can easily deduce a relation between ρ_1 and $\rho_2 \neq \rho_1$ where equation (D.10) still holds and the expression in (D.9) is still negative. We prove in Appendix E, that the function $F(k,\rho)$ is decreasing in $\rho \in [0,1]$. Consequently, this proves that the expression in (D.1) is concave when $\rho_1 \leq \rho_2$.

On the other hand, the results of the numerical experiment shown in Figure 4.1 suggests some (ρ_1,ρ_2) pairs such that $\rho_2 \leq \rho_1$ might also be extremal. In the sequel, we disprove that the expression in (D.1) is concave when $\rho_1 > \rho_2$ providing a counter example. Hence, the extremality in the case $\rho_1 > \rho_2$ is not based on the concavity of (D.1).

Let us try to analyze the limiting case when $k = 1$.

$$\lim_{k\to 1} \Phi(k, \rho_1, \rho_2) = \frac{(1+\rho_1)\,\rho_2}{\rho_1\,(1+\rho_2)}$$
$$\lim_{k\to 1} \Psi(k, \rho_2) = \frac{1}{\rho_2\,(1+\rho_2)}$$
$$\lim_{k\to 1} \frac{\Psi'(k,\rho_2)}{\Psi(k,\rho_2)} = \lim_{k\to 1} \frac{\dfrac{(1+\rho_2)}{\rho_2} - \dfrac{\left(k^{\frac{\rho_2}{1+\rho_2}} + k^{\frac{1}{1+\rho_2}}\right)\log k}{\gamma(k,\rho_2)}}{(1+\rho_2)^2 \log\left(\frac{1}{2} + \frac{1}{2}k^{\frac{1}{1+\rho_2}}\right) + \dfrac{\left(1 + k^{\frac{1}{1+\rho_2}} - \rho_2\left(k - k^{\frac{1}{1+\rho_2}}\right)\right)\log k}{\gamma(k,\rho_2)}} = 0$$
$$\lim_{k\to 1} F(k,\rho_2) - F(k,\rho_1) = \frac{\rho_2}{(1+\rho_2)\left(k^{\frac{\rho_2}{1+\rho_2}} - 1\right)} - \frac{\rho_1}{(1+\rho_1)\left(k^{\frac{\rho_1}{1+\rho_1}} - 1\right)} = \infty - \infty$$

To simplify notation we define $s_1 = \frac{\rho_1}{1+\rho_1}$ and $s_2 = \frac{\rho_2}{1+\rho_2}$.

$$\lim_{k\to 1} \frac{s_2}{k^{s_2}-1} - \frac{s_1}{k^{s_1}-1} = \lim_{k\to 1} \frac{s_2\,(k^{s_1}-1) - s_1\,(k^{s_2}-1)}{(k^{s_1}-1)\,(k^{s_1}-1)} = \frac{0}{0}$$
$$\lim_{k\to 1} \frac{s_1 s_2\,(k^{s_1} - k^{s_2})}{s_2 k^{s_2}\,(k^{s_1}-1) + s_1 k^{s_1}\,(k^{s_2}-1)} = \frac{0}{0}$$
$$\lim_{k\to 1} \frac{s_1 s_2\,(s_1 k^{s_1} - s_2 k^{s_2})}{k^{s_1+s_2}\,(s_1+s_2)^2 - s_1^2 k^{s_1} - s_2^2 k^{s_2}} = \frac{s_1 - s_2}{2}$$

Therefore using the product rule of limits, we conclude that

$$\begin{aligned}&\lim_{k\to 1} \Phi(k,\rho_1,\rho_2)\Psi(k,\rho_2)\left(\frac{\Psi'(k,\rho_2)}{\Psi(k,\rho_2)} + F(k,\rho_2) - F(k,\rho_1)\right)\\ =&2^{-1}\frac{(1+\rho_1)}{\rho_1\,(1+\rho_2)^2}\left(\frac{\rho_1}{1+\rho_1} - \frac{\rho_2}{1+\rho_2}\right)\end{aligned}$$

The fact that $\rho/(1+\rho)$ is increasing in ρ implies that the next inequalities hold

$$\lim_{k\to 1} \Phi(k,\rho_1,\rho_2)\Psi(k,\rho_2)\left(\frac{\Psi'(k,\rho_2)}{\Psi(k,\rho_2)} + F(k,\rho_2) - F(k,\rho_1)\right) \geq 0 \quad \text{for} \quad \rho_1 > \rho_2$$
$$\lim_{k\to 1} \Phi(k,\rho_1,\rho_2)\Psi(k,\rho_2)\left(\frac{\Psi'(k,\rho_2)}{\Psi(k,\rho_2)} + F(k,\rho_2) - F(k,\rho_1)\right) \leq 0 \quad \text{for} \quad \rho_1 \leq \rho_2$$

As a consequence, we know that when $k \to 1$ the expression in (D.9) is always positive for $\rho_1 > \rho_2$. This in turn implies that the concavity region of (D.1) is limited to $\rho_1 \leq \rho_2$ values.

E

Appendix E

In this appendix, we first prove that the function $F(k,\rho)$ defined in (D.6) is a decreasing function in $\rho \in [0,1]$. Then, we show the inequality given in (D.8) holds.

For convenience, we define the function $H(k,\rho) = -\frac{k}{1+k}F(k,\rho)$ as

$$H(k,\rho) = \frac{\rho}{1+\rho}\frac{1}{\left(1+k^{\frac{1}{1+\rho}}\right)\left(1-k^{\frac{\rho}{1+\rho}}\right)} \geq 0 \tag{E.1}$$

where $k \in [0,1]$. We note that instead of $F(k,\rho)$, we can also check the monotonicity of $H(k,\rho)$ with respect to ρ.

We now follow a series of transformations. Let

$$t = \frac{\rho}{1+\rho} \quad \text{for} \quad t \in [0, \frac{1}{2}]$$

Then, (E.1) reduces to

$$H(k, \frac{t}{1-t}) = \frac{t}{(1-k^t)\,(1+k^{1-t})}$$

In addition, let

$$s = -t\ln k \quad \text{for} \quad s \in [0, \frac{1}{2}\ln\frac{1}{k}]$$

Then,

$$H(k, \frac{-s}{\log k + s}) = \frac{1}{\log\frac{1}{k}} \;\; \frac{s}{1-e^{-s}} \;\; \frac{1}{1+ke^{s}} \tag{E.2}$$

We note that the first fraction in (E.2) can be treated as a constant and we ignore it. We define the variable $a = \frac{1}{k} \geq 1$. For simplicity, we consider the function

$$\frac{1}{H(k, \frac{-s}{\log k+s})} = \underbrace{\frac{\ln a}{a}}_{constant} \;\; \frac{1-e^{-s}}{s}\,(a+e^{s})$$

We first show that $\ln\left(\frac{1-e^{-s}}{s}(a+e^s)\right)$ is a convex function for all $s \geq 0$. Taking the first derivative with respect to s, we obtain

$$\frac{\partial}{\partial s}\left(-\ln s + \ln\left(\frac{1}{1-e^{-s}}\right) + \ln\left(\frac{e^s}{a+e^s}\right)\right) = -\frac{1}{s} + \frac{e^s}{a+e^s} + \frac{1}{e^s-1} \tag{E.3}$$

Taking the second derivative in s, we get

$$\begin{aligned}
&\frac{\partial^2}{\partial s^2}\left(-\ln s + \ln\left(\frac{1}{1-e^{-s}}\right) + \ln\left(\frac{e^s}{a+e^s}\right)\right) \\
=\ &\frac{1}{s^2} + \frac{ae^s}{(a+e^s)^2} - \frac{e^s}{(e^s-1)^2} \\
\geq\ &\frac{1}{s^2} - \frac{e^s}{(e^s-1)^2} \\
=\ &\frac{1}{s^2} - \left(\frac{1}{e^{\frac{s}{2}} + e^{\frac{-s}{2}}}\right)^2 \\
=\ &\frac{1}{s^2} - \frac{1}{\left(2\sinh\frac{s}{2}\right)^2} \\
\geq\ &0
\end{aligned}$$

where the last inequality follows from $\sinh x$ Taylor expansion

$$\sinh x = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} \geq x$$

We proved that $\ln\left(\frac{1-e^{-s}}{s}(a+e^s)\right)$ is a convex function for all $s \geq 0$. Therefore the function has only one minimum, and to decide whether the expression is decreasing in $s \in [0, \frac{1}{2}\ln a]$, it is sufficient to evaluate (E.3) at $s = \frac{1}{2}\ln a$.

$$\begin{aligned}
&\frac{\partial}{\partial s}\left(-\ln s + \ln\left(\frac{1}{1-e^{-s}}\right) + \ln\left(\frac{e^s}{a+e^s}\right)\right)\Big|_{s=\frac{1}{2}\ln a} \\
=&-\frac{1}{\ln\sqrt{a}} + \frac{\sqrt{a}}{a+\sqrt{a}} + \frac{1}{\sqrt{a}-1} \\
=&-\frac{1}{\ln\sqrt{a}} + \frac{2\sqrt{a}}{a-1} \\
\leq\ &0
\end{aligned}$$

since for $b = \sqrt{a} \geq 1$, we can show that

$$\frac{b^2-1}{2b} - \ln b \geq 0 \tag{E.4}$$

Taking the first derivative of (E.4) with respect to b, we get

$$\frac{\partial}{\partial b}\frac{b^2-1}{2b} - \ln b = \frac{1}{2} + \frac{1}{2b^2} - \frac{1}{b} = \frac{(b-1)^2}{2b^2} \geq 0$$

Therefore, we proved that for each $k \in [0,1]$ the function $\frac{1}{H(k,\frac{-s}{\log k+s})}$ is decreasing in s. By definition, the variable t is increasing in ρ, and $s = -t \ln k$ is also increasing in t for a given k. As a consequence, the function $F(k,\rho) = -\dfrac{1+k}{k} H(k,\rho)$ is decreasing in ρ.

We now show that the inequality given in (D.8) holds. Using the above derivations, this is equivalent to the following:

$$-H(k, \frac{-s}{\log k + s}) \left(e^{-s} + ke^{s}\right) \log k \leq 1.$$

Using the expression in (E.2), we get

$$\left(e^{-s} + ke^{s}\right) \log k \geq -\frac{1}{H(k,\frac{-s}{\log k+s})} = \frac{1-e^{-s}}{s} \left(1 + ke^{s}\right) \log k \tag{E.5}$$

Simplifying the $\log k$ in both sides, we only need to show that for $s \in [0, \frac{1}{2} \ln \frac{1}{k}]$

$$s\left(e^{-s} + ke^{s}\right) \leq \left(1 - e^{-s}\right)\left(1 + ke^{s}\right)$$

holds. First, we note that we have equality when $s = 0$. Moreover, taking the derivative of both sides with respect to s, we have

$$\frac{\partial}{\partial s}\left[s\left(e^{-s} + ke^{s}\right)\right] = \left(e^{-s} + ke^{s}\right) + s\left(ke^{s} - e^{-s}\right)$$
$$\frac{\partial}{\partial s}\left[\left(1 - e^{-s}\right)\left(1 + ke^{s}\right)\right] = \left(e^{-s} + ke^{s}\right)$$

Therefore, for any $k \in [0,1]$, and $s \in [0, \frac{1}{2} \ln \frac{1}{k}]$

$$\frac{\partial}{\partial s}\left[s\left(e^{-s} + ke^{s}\right)\right] \leq \frac{\partial}{\partial s}\left[\left(1 - e^{-s}\right)\left(1 + ke^{s}\right)\right]$$

since

$$s\left(ke^{s} - e^{-s}\right) = se^{-s}(ke^{2s} - 1) \leq 0.$$

as $ke^{2s} = k^{\frac{1-\rho}{1+\rho}} \leq 1$. This concludes the proof.

F

Appendix F

We define the function $H_{z,\rho}(k) : [2^{-\rho}, 1] \to [2^{-\rho}, g(\rho, z)]$ as

$$H_{z,\rho}(k) = h(\rho, g^{-1}(\rho, k), z) \tag{F.1}$$

where

$$g(\rho, z) \triangleq \left(\frac{1}{2}(1+z)^{\frac{1}{1+\rho}} + \frac{1}{2}(1-z)^{\frac{1}{1+\rho}}\right)^{1+\rho}$$
$$h(\rho, z_1, z_2) \triangleq \frac{1}{2}(1+z_1z_2)g(\rho, \frac{z_1+z_2}{1+z_1z_2}) + \frac{1}{2}(1-z_1z_2)g(\rho, \frac{z_1-z_2}{1-z_1z_2}).$$

In this Appendix, we show that the function $H_{z,\rho}(k)$ is concave with respect to the variable $k \in [0, 1]$ for fixed values of $z, \rho \in [0, 1]$.

Taking the first derivative, we get

$$\frac{\partial}{\partial\, t} H_{z,\rho}(k) = \frac{h'(\rho, g^{-1}(\rho, k), z)}{g'(\rho, g^{-1}(\rho, k))}.$$

As done in Appendix A, we define $u = g^{-1}(\rho_1, k)$. Since $g(\rho, u)$ is a decreasing function in u, so is $g^{-1}(\rho_1, k)$ in k. Hence we can check the concavity of $H_{z,\rho_1}(k)$ with respect to variable k, by verifying that

$$\frac{h'(\rho, u, z)}{g'(\rho, u)}.$$

is increasing in u. Hence, we check that

$$\frac{\partial}{\partial u}\left(\frac{h'(\rho, u, z)}{g'(\rho, u)}\right) = \frac{h''(\rho, u, z)g'(\rho, u) - h'(\rho, u, z)g''(\rho, u)}{g'(\rho, u)^2} \geq 0.$$

Since the denominator is always positive, we only need to show that

$$h''(\rho, u, z)g'(\rho, u) - h'(\rho, u, z)g''(\rho, u) \geq 0. \tag{F.2}$$

Moreover, we observe that $h(\rho, u, 0) = g(\rho, u)$. So, we can equivalently show the following relation holds:

$$\frac{h''(\rho, u, z)}{h'(\rho, u, z)} \geq \frac{h''(\rho, u, 0)}{h'(\rho, u, 0)}. \tag{F.3}$$

We first apply the transformations

$$u = \tanh(t), \qquad z = \tanh(w)$$

where $t, w \in [0, \infty)$. For shorthand notation, let $h(\rho, \tanh(t), \tanh(w)) \triangleq \tilde{h}(\rho, t, w)$. Using these, we obtain

$$\tilde{h}(\rho, t, w) = \frac{\cosh(\frac{1}{1+\rho}(t+w))^{1+\rho} + \cosh(\frac{1}{1+\rho}(t-w))^{1+\rho}}{2\cosh(t)\cosh(w)}.$$

Then,

$$\frac{\frac{\partial h(\rho,t,w)}{\partial t^2}}{\frac{\partial h(\rho,t,w)}{\partial t}} = -2\tanh(t) + \frac{\rho}{1+\rho}\cosh(t)\times$$
$$\left[\frac{\cosh(\frac{1}{1+\rho}(t+w))^{\rho-1} + \cosh(\frac{1}{1+\rho}(t-w))^{\rho-1}}{\cosh(\frac{1}{1+\rho}(t+w))^{\rho}\sinh(\frac{\rho}{1+\rho}t - \frac{1}{1+\rho}w) + \cosh(\frac{1}{1+\rho}(t-w))^{\rho}\sinh(\frac{\rho}{1+\rho}t + \frac{1}{1+\rho}w)}\right]. \tag{F.4}$$

We note that the additive term $-2\tanh(t)$, and the non-negative multiplicative factor $\frac{\rho}{1+\rho}\cosh(t)$ do not depend on w. Hence, we only need to show the term inside the paranthesis is smallest when evaluated at $w = 0$. For this purpose, we define the transformations

$$a = \frac{t+w}{1+\rho}, \qquad b = \frac{t-w}{1+\rho}$$

such that $t = (1+\rho)\frac{a+b}{2}$, and $w = (1+\rho)\frac{a-b}{2}$. The condition $t, w \geq 0$ is equivalent to $a \geq |b|$. Using these transformations, the reciprocal of the term inside paranthesis in equation (F.4) becomes

$$R(\rho, a, b) = \frac{\cosh(b)^{1-\rho}\cosh(a)\sinh(\frac{a+b}{2}\rho - \frac{a-b}{2}) + \cosh(a)^{1-\rho}\cosh(b)\sinh(\frac{a+b}{2}\rho + \frac{a-b}{2})}{\cosh(a)^{1-\rho} + \cosh(b)^{1-\rho}}.$$

Therefore, the inequality given in (F.3) will hold iff

$$R(\rho, a, b) \leq R(\rho, \frac{a+b}{2}, \frac{a+b}{2}) = \cosh(\frac{a+b}{2})\sinh(\frac{a+b}{2}\rho). \tag{F.5}$$

We define

$$f(\rho, a, b) \triangleq \cosh(\frac{a+b}{2})\sinh(\rho\frac{a+b}{2})\left[\cosh(a)^{1-\rho} + \cosh(b)^{1-\rho}\right]$$
$$-\cosh(a)^{1-\rho}\cosh(b)\sinh(\rho\frac{a+b}{2} + \frac{a-b}{2}) - \cosh(b)^{1-\rho}\cosh(a)\sinh(\rho\frac{a+b}{2} - \frac{a-b}{2}).$$

We note that $f(\rho, a, b) \geq 0$ is equivalent to the inequality (F.5), which in turn is equivalent to the inequality (F.3).

After simplifications, the function reduces to the following form:

$$f(\rho, a, b) = \sinh(\frac{a-b}{2}) J(\rho, a, b)$$

where

$$J(\rho, a, b) \triangleq \cosh(b)^{1-\rho} \cosh(a - \rho\frac{a+b}{2}) - \cosh(a)^{1-\rho} \cosh(b - \rho\frac{a+b}{2}).$$

Since for $a \geq |b|$, we have

$$\sinh(\frac{a-b}{2}) \geq 0$$

we only need to show that $J(\rho, a, b) \geq 0$.

We introduce the variables t', and w' using $a = t' + w'$, and $b = t' - w'$ where $t', w' \in [0, \infty)$. Then, we get

$$J(\rho, t'+w', t'-w') = \cosh(t'-w')^{1-\rho} \cosh(t'-\rho t'+w') - \cosh(t'-\rho t'-w') \cosh(t'+w')^{1-\rho}.$$

We note that $J(\rho, t' + w', t' - w')\Big|_{t'=0} = 0$. Moreover, $J(\rho, t' + w', t' - w')$ is increasing in the variable t': taking the first derivative with respect to t', we get

$$\frac{\partial}{\partial t'} J(\rho, t'+w', t'-w') = (1-\rho) \left[\cosh(t' - w')^{-\rho} - \cosh(t' + w')^{-\rho}\right] \sinh((2-\rho)t') \geq 0$$

where the positivity follows from the fact that $|t' - w'| \leq |t' - w'|$, thus $\cosh(t' - w') \leq \cosh(t' + w')$, and $\cosh(t' - w')^{-\rho} \geq \cosh(t' + w')^{-\rho}$, and from the fact that $\sinh(x) \geq 0$ holds for $\forall x \geq 0$.

As a result, $J(\rho, t' + w', t' - w') \geq 0$ as required, and we have shown that the inequality given in (F.3) holds. This concludes the proof.

Bibliography

[1] E. Arıkan. Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Trans. Inf. Theor.*, 55(7):3051–3073, 2009.

[2] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, USA, 1968.

[3] E. Arıkan and E. Telatar. BEC and BSC are E_0 extremal. Unpublished.

[4] E. Arıkan and E. Telatar. On the rate of channel polarization. abs/0807.3806, 2008.

[5] E. Arıkan. A performance comparison of polar codes and Reed-Muller codes. *IEEE Commun. Lett.*, 12(6):447–449, June 2008.

[6] S. H. Hassani, S. B. Korada, and R. L. Urbanke. The compound capacity of polar codes. abs/0907.3291, 2009.

[7] E. Arıkan. An inequality on guessing and its application to sequential decoding. *IEEE Transactions on Information Theory*, 42(1):99–105, 1996.

[8] A. Rényi. On measures of entropy and information. *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, Vol. 1:547–561, 1961.

[9] S. Arimoto. Information measures and capacity of order α for discrete memoryless channels. In *Topics in information theory*, I.Csiszar and P. Elias, editors, Amsterdam, The Netherlands, 1977. North-Holland Publishing Co.

[10] J. L. Massey. Guessing and entropy. In *Proceedings of the 1994 IEEE International Symposium on Information Theory*, page 204, 1994.

[11] D. Blackwell, L. Breiman, and A. J. Thomasian. The capacity of a class of channels. *The Annals of Mathematical Statistics*, 3(4):12291241, 1959.

[12] Satish Babu Korada. *Polar codes for channel and source coding*. PhD thesis, Lausanne, 2009.

[13] E. Şaşoğlu. Private communication, October 2009.

[14] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series and Products*. Academic Press Inc, 1994.