Upper Bounds to the Correct-Decoding Probability under Minimum Likelihood Decoding

Alfonso Martinez
Universitat Pompeu Fabra
alfonso.martinez@ieee.org

Josep Font-Segura
Universitat Pompeu Fabra
josep.font@ieee.org

Albert Guillén i Fàbregas
University of Cambridge
Universitat Politècnica de Catalunya
guillen@ieee.org

***Abstract*—While the correct-decoding probability under maximum likelihood decoding is lower bounded by the inverse of the number of messages, i.e. an exponential rate of decay equal to the rate R, much lower correct-decoding probabilities are possible under minimum likelihood decoding, the worst possible decoding rule. This paper provides an upper bound to the exponential rate of decay of the lowest possible correct-decoding probability under minimum likelihood decoding. The upper bound is characterized by an optimization problem involving a Rényi divergence of negative order, or equivalently Gallager's channel-coding function $E_0(\rho)$ with ρ below -1, thus providing an operational meaning to either of them.**

***Index Terms*—Minimum likelihood decoding, unreliability function, Rényi divergence, Gallager's E_0 function, metaconverse.**

I. Introduction

While the pursuit of reliable communication has driven decades of research in coding theory, an equally fascinating but underexplored frontier exists at the opposite end of the spectrum: deliberately bad codes. In addition to its intrinsic interest, the study of bad codes has implications for adversarial communications, where sophisticated attackers might manipulate the decoder internal operation to maximally disrupt the transmission of information. Understanding the theoretical limits of such attacks to the decoder in the receiver, rather than to the channel, provides essential insights into fundamental security bounds for systems operating in hostile environments.

In adversarial communication, bad codes should be decoded so as to minimize the probability of correct decoding, as we review next. Let $P_M(m)$ denote the probability of sending message m, $\boldsymbol{x}_m$ the corresponding codeword (of length n), $\mathcal{C}$ the code of rate R and $M_n = 2^{nR}$ codewords (and messages), $\boldsymbol{y}$ the channel output (also of length n), $W^n(\boldsymbol{y}|\boldsymbol{x})$ the channel transition probability, and $\hat{m}(\boldsymbol{y})$ the decoder output. The decoder may be randomized, e.g. in the presence of ties, with decoder output described by the probability $P_{\hat{M}|\boldsymbol{Y}}(\hat{m}|\boldsymbol{y})$; $\hat{m}(\boldsymbol{y})$ is always a message, for the error probability might otherwise easily be made equal to one. The probability of correct decoding $P_{\mathrm{c}}(\mathcal{C})$ is then given by

$$P_{\mathrm{c}}(\mathcal{C}) = \sum_{m,\boldsymbol{y},\hat{m}:m=\hat{m}} P_M(m) W^n(\boldsymbol{y}|\boldsymbol{x}_m) P_{\hat{M}|\boldsymbol{Y}}(\hat{m}|\boldsymbol{y}). \quad (1)$$

This work has been funded in part by the Spanish Ministry of Science, Innovation, and Universities under grant PID2024-159557OB-C22 and by the European Research Council under grant 101142747.

For a given $\boldsymbol{y}$, the probability $P_{\mathrm{c}}(\mathcal{C})$ is lowest if $\hat{m}$ is chosen such that $P_M(\hat{m})W^n(\boldsymbol{y}|\boldsymbol{x}_{\hat{m}}) = P_{\boldsymbol{Y}}(\boldsymbol{y})P_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}_{\hat{m}}|\boldsymbol{y})$ is smallest. If there is only one such message, then $P_{\hat{M}|\boldsymbol{Y}}(\hat{m}|\boldsymbol{y}) = 1$. The ways ties are resolved if there are several such messages does not affect the overall probability $P_{\mathrm{c}}(\mathcal{C})$. This extreme form of decoding is minimum a posteriori and degenerates to minimum likelihood (MinL) decoding when all messages are equiprobable. Under MinL, the probability of correct decoding in (1), denoted by $\underline{P_{\mathrm{c}}}(\mathcal{C})$, can thus be expressed as

$$\underline{P_{\mathrm{c}}}(\mathcal{C}) = \frac{1}{M_n} \sum_{\boldsymbol{y}} \min_m W^n(\boldsymbol{y}|\boldsymbol{x}_m). \quad (2)$$

By construction, MinL decoding satisfies $\underline{P_{\mathrm{c}}}(\mathcal{C}) \leq \frac{1}{M_n}$ and a decoding error probability $\underline{P_{\mathrm{e}}}(\mathcal{C}) \geq 1 - \frac{1}{M_n}$, with equality occurring when all codewords are identical or when the channel is extremely noisy. Eq. (2) is similar to that of maximum likelihood (MaxL) decoding, whose correct-decoding probability $\overline{P_{\mathrm{c}}}(\mathcal{C})$ is computed by replacing $\min$ with $\max$ in the equation above. In contrast to MinL, MaxL decoding satisfies $\overline{P_{\mathrm{c}}}(\mathcal{C}) \geq \frac{1}{M_n}$ and $\overline{P_{\mathrm{e}}}(\mathcal{C}) \leq 1 - \frac{1}{M_n}$. When studying bad codes, we wish to minimize $\underline{P_{\mathrm{c}}}(\mathcal{C})$, pushing down the probability as far from $1/M_n$ as possible.

Since reliable communication is impossible with bad codes in adversarial scenarios, we investigate a new metric: the channel unreliability function $\underline{E}(R)$. This function quantifies the exponential decay (in the blocklength n) of the correct-decoding probability of the worst possible codes of rate R:

$$\underline{E}(R) = \limsup_{n\to\infty} \left\{ -\frac{1}{n} \log \underline{P_{\mathrm{c}}^*}(n, 2^{nR}) \right\}, \quad (3)$$

where $\underline{P_{\mathrm{c}}^*}(n, M_n)$ denotes the lowest possible correct-decoding probability for given n and M_n,

$$\underline{P_{\mathrm{c}}^*}(n, M_n) = \min_{\mathcal{C}} \underline{P_{\mathrm{c}}}(\mathcal{C}), \quad (4)$$

where the minimum is over all codes with M_n codewords of length n.

In this paper we determine an upper bound to the unreliability function $\underline{E}(R)$ for discrete memoryless channels:

$$\underline{E}(R) \leq \min_{\rho<-1} \max_{Q} \big\{ E_0(\rho, Q) - \rho R \big\}, \quad (5)$$

where $E_0(\rho, Q)$ is Gallager's channel-coding function in [1, Eq. (5.6.14)] for a given single-letter input probability distribution $Q(x)$. This analysis gives operational meaning to Gallager's $E_0(\rho)$ function for $\rho < -1$.

A. Single-letter Gallager's channel-coding function

For discrete-memoryless channels with input probability distribution $Q(x)$ and channel transition probability $W(y|x)$ (or $W_x(y)$) and $\rho \neq -1$, the channel-coding functions $G_{\rho,Q}(y)$, $F_0(\rho, Q)$, and $E_0(\rho, Q)$ are respectively given by

$$G_{\rho,Q}(y) = \left(\sum_x Q(x) W(y|x)^{\frac{1}{1+\rho}}\right)^{1+\rho}, \tag{6}$$

$$F_0(\rho, Q) = \sum_y G_{\rho,Q}(y), \tag{7}$$

$$E_0(\rho, Q) = -\log F_0(\rho, Q). \tag{8}$$

Evaluating the Gallager function for $\rho < -1$ requires some care if there is some x for which $Q(x) > 0$ and $W(y|x) = 0$. In this case we use a regularization procedure: replace each $W(y|x)$ by an infinitesimal quantity $W(y|x) = \delta_{y,x} > 0$, evaluate the expression, and then take the limit $\delta_{y,x} \to 0$. Now, for $\rho < -1$, the term $W(y|x)^{\frac{1}{1+\rho}}$ diverges to infinity as $\delta_{y,x} \to 0$, dominating the summation over x. After raising the sum to the power $1+\rho$, we obtain a quantity proportional to the infinitesimal $\delta_{y,x}$, and $G_{\rho,Q}(y)$ vanishes in the limit $\delta_{y,x} \to 0$ when there exists an x such that $Q(x) > 0$ and $W(y|x) = 0$. The only non-zero contributions in (7) come from values of y such that $W(y|x) > 0$ for all x for which $Q(x) > 0$. For example, the function $F_0(\rho, Q)$ for the binary-erasure channel with erasure probability ε and $q = Q(0)$ is given by

$$F_0(\rho, Q) = \begin{cases} \big(q^{1+\rho} + (1-q)^{1+\rho}\big)(1-\varepsilon) + \varepsilon, & \rho > -1, \\ \varepsilon, & \rho < -1. \end{cases} \tag{9}$$

B. Connections with Rényi divergence

The function $G_{\rho,Q}(y)$ induces a tilted output distribution, $R_Y^{\rho,Q}(y)$, given by

$$R_Y^{\rho,Q}(y) = \frac{G_{\rho,Q}(y)}{F_0(\rho, Q)}. \tag{10}$$

The channel coding function $E_0(\rho, Q)$ can be expressed in terms of the Rényi divergence $D_{\frac{1}{1+\rho}}\big(QW \,\|\, QR_Y^{\rho,Q}\big)$ of order $\frac{1}{1+\rho}$ between the distributions QW and $QR_Y^{\rho,Q}$ as

$$E_0(\rho, Q) = \rho D_{\frac{1}{1+\rho}}\big(QW \,\|\, QR_Y^{\rho,Q}\big), \tag{11}$$

where the Rényi divergence $D_\alpha(A \,\|\, B)$ of order α between distributions $A(x, y)$ and $B(x, y)$ is given by [2, Eq. (3.3)]

$$D_\alpha(A \,\|\, B) = \frac{1}{\alpha - 1} \log\left(\sum_{x,y} A^\alpha(x, y) B^{1-\alpha}(x, y)\right). \tag{12}$$

C. Multi-letter Gallager's functions

Given a code $\mathcal{C}$ with distribution P_M over messages m, or equivalently over codewords $\boldsymbol{x}_m$, channel transition probability W^n, and $\rho \neq -1$, we define a multi-letter channel-coding function $F_0(\rho, P_M)$ as

$$F_0(\rho, P_M) = \sum_{\boldsymbol{y}} \left(\sum_m P_M(m) W^n(\boldsymbol{y}|\boldsymbol{x}_m)^{\frac{1}{1+\rho}}\right)^{1+\rho}, \tag{13}$$

and $E_0(\rho, P_M) = -\log F_0(\rho, P_M)$. The tilted output distribution $R_{\boldsymbol{Y}}^{\rho,P_M}(\boldsymbol{y})$ is given by

$$R_{\boldsymbol{Y}}^{\rho,P_M}(\boldsymbol{y}) = \frac{\left(\sum_m P_M(m) W^n(\boldsymbol{y}|\boldsymbol{x}_m)^{\frac{1}{1+\rho}}\right)^{1+\rho}}{F_0(\rho, P_M)}. \tag{14}$$

As $\rho \to -1^-$, that is the Rényi divergence of order $-\infty$, the distribution $R_{\boldsymbol{Y}}^{\rho,P_M}(\boldsymbol{y})$ converges to the probability distribution induced by MinL decoding, cf. (2):

$$\underline{R}_{\boldsymbol{Y}}^*(\boldsymbol{y}) = \frac{1}{\mu(\mathcal{C})} \min_m W^n(\boldsymbol{y}|\boldsymbol{x}_m), \tag{15}$$

where $\mu(\mathcal{C}) = \sum_{\boldsymbol{y}} \min_m W^n(\boldsymbol{y}|\boldsymbol{x}_m) = M_n \underline{P}_{\mathrm{c}}(\mathcal{C})$. This limit provides a first concrete link between MinL decoding, negative-order Rényi divergences.

D. Paper organization

For maximum likelihood decoding, recent work by Polyanskiy, Poor, and Verdú showed that a number of classical converse bounds can be obtained in a unified manner from an umbrella metaconverse bound [3, Th. 30], that lower bounds the error probability that all codes must satisfy in the range $0 \leq P_{\mathrm{e}} \leq 1 - \frac{1}{M_n}$. The metaconverse bound is derived from a properly chosen binary hypothesis test between two alternatives induced by $\mathcal{C}$. The distribution $\underline{R}_{\boldsymbol{Y}}^*$ in Eq. (15) is instrumental in relating MinL decoding to optimal testing between two hypotheses $A(m, \boldsymbol{y})$ and $B(m, \boldsymbol{y})$ induced by the code for the generation of an observation $(m, \boldsymbol{y})$:

$$A(m, \boldsymbol{y}) = P_M(m) W^n(\boldsymbol{y}|\boldsymbol{x}_m), \tag{16}$$

$$B(m, \boldsymbol{y}) = P_M(m) \underline{R}_{\boldsymbol{Y}}^*(\boldsymbol{y}). \tag{17}$$

We start in Sect. II with a brief review of binary hypothesis testing, with particular emphasis on its extremal tests, the standard Neyman-Pearson test which minimizes crossover probability and the less commonly studied anti-Neyman-Pearson test which maximizes it. Both tests will be essential to our characterization of the channel unreliability function.

Then, in Sect. III, we extend the metaconverse bound to minimum likelihood decoding to obtain a fundamental lower limit on the correct-decoding probability that all codes must satisfy under MinL decoding, which corresponds to the range $1 - \frac{1}{M_n} \leq P_{\mathrm{e}} \leq 1$. In contrast to the metaconverse for maximum likelihood decoding, that yields a lower limit on the error probability, the modified metaconverse characterizes the worst possible performance rather than the best possible performance, and considers the worst possible Neyman-Pearson test, rather than the best one.

We conclude in Sect. V with a derivation of an upper bound on the exponent of the metaconverse, which yields (5). This derivation follows similar lines to those of Polyanskiy and Verdú on the exponent of Arimoto's strong converse in [4].

II. Extremal tests in binary hypothesis testing

Binary hypothesis testing is the problem of choosing between two alternative probability distributions $A(x)$ and $B(x)$ for generating an observation x. A (randomized) test $T(x)$

assigns one of the two probability distributions, $T(x) \in \{A, B\}$, to each observation x. Every test T has two associated cross-over probabilities $\varepsilon_{A|B}(T)$ and $\varepsilon_{B|A}(T)$, namely the probability of choosing A (resp. B) when x is generated according to B (resp. A):

$$\varepsilon_{A|B}(T) = \sum_x B(x) \Pr\{T(x) = A\}, \tag{18}$$

$$\varepsilon_{B|A}(T) = \sum_x A(x) \Pr\{T(x) = B\}. \tag{19}$$

Of special interest are the extreme tests which attain the lowest or highest value of one cross-over probability, say $\varepsilon_{A|B}(T)$, having fixed the other one, $\varepsilon_{B|A}(T) = \alpha$. Let $\underline{\beta}_\alpha(A, B)$ (resp. $\overline{\beta}_\alpha(A, B)$) denote the lowest (resp. highest) $\varepsilon_{A|B}(T)$ for fixed $\varepsilon_{B|A}(T) = \alpha$, that is

$$\underline{\beta}_\alpha(A, B) = \min_{T:\varepsilon_{B|A}(T)\leq\alpha} \varepsilon_{A|B}(T), \tag{20}$$

$$\overline{\beta}_\alpha(A, B) = \max_{T:\varepsilon_{B|A}(T)\geq\alpha} \varepsilon_{A|B}(T). \tag{21}$$

Both expressions are related by $\underline{\beta}_\alpha(A, B) + \overline{\beta}_{1-\alpha}(A, B) = 1$. The inverse functions $\underline{\alpha}_\beta(A, B)$ and $\overline{\alpha}_\beta(A, B)$ that give the extreme values of $\varepsilon_{B|A}(T)$ for fixed $\varepsilon_{A|B}(T)$ are defined analogously. Extremal tests are likelihood ratio tests with $\varepsilon_{B|A}(T) = \alpha$. The minimizing test is the usual Neyman-Pearson test; we refer to the maximizing test as the anti-Neyman-Pearson test.

III. A metaconverse for Minimum Likelihood decoding

We now establish a metaconverse bound for minimum likelihood decoding that provides a fundamental lower limit on the correct-decoding probability, adapting the metaconverse for maximum likelihood to characterize worst-case rather than best-case performance.

Let $\hat{m}(\boldsymbol{y})$ be the output of a MinL decoder; if there are ties, the decoder selects uniformly at random from $\mathcal{S}(m, \boldsymbol{y})$, the set of messages $\overline{m}$ such that $W^n(\boldsymbol{y}|\boldsymbol{x}_{\overline{m}}) = W^n(\boldsymbol{y}|\boldsymbol{x}_m)$. We consider two tests in our analysis. The first test, T_{MinL}, is based directly on the MinL decoder: if $m = \hat{m}(\boldsymbol{y})$, select hypothesis $A = P_M W^n$; otherwise, $B = P_M \underline{R}^*_{\boldsymbol{Y}}$ is selected.

The second test, T^*_{MinL}, is an anti-Neyman-Pearson test with $\lambda = \mu(\mathcal{C})^{-1}$. This test selects $P_M W^n$ if $\mu(\mathcal{C}) P_M(m) \underline{R}^*_{\boldsymbol{Y}}(\boldsymbol{y}) > P_M(m) W^n(\boldsymbol{y}|\boldsymbol{x}_m)$, that is, if

$$\min_{\overline{m}} W^n(\boldsymbol{y}|\boldsymbol{x}_{\overline{m}}) > W^n(\boldsymbol{y}|\boldsymbol{x}_m), \tag{22}$$

an inequality that clearly never holds. Also, the test T^*_{MinL} selects $P_M W^n$ with probability $\pi_A(m, \boldsymbol{y})$, the inverse of the cardinality of the set $\mathcal{S}(m, \boldsymbol{y})$ of messages defined earlier, if

$$\min_{\overline{m}} W^n(\boldsymbol{y}|\boldsymbol{x}_{\overline{m}}) = W^n(\boldsymbol{y}|\boldsymbol{x}_m), \tag{23}$$

and $P_M \underline{R}^*_{\boldsymbol{Y}}$ with probability $1 - \pi_A(m, y)$ if the above condition holds. Finally, the test T^*_{MinL} also selects $P_M \underline{R}^*_{\boldsymbol{Y}}$ with probability 1 when

$$\min_{\overline{m}} W^n(\boldsymbol{y}|\boldsymbol{x}_{\overline{m}}) < W^n(\boldsymbol{y}|\boldsymbol{x}_m). \tag{24}$$

Since both tests T_{MinL} and T^*_{MinL} make the same decisions, $\varepsilon_{A|B}$ and $\varepsilon_{B|A}$ can be respectively computed as

$$\varepsilon_{A|B}(T_{\mathrm{MinL}}) = \sum_{m,\boldsymbol{y}} P_M(m) \underline{R}^*_{\boldsymbol{Y}}(\boldsymbol{y}) \mathbf{1}\{m = \hat{m}(\boldsymbol{y})\} \tag{25}$$

$$= \frac{1}{M_n} \tag{26}$$

and

$$\varepsilon_{B|A}(T_{\mathrm{MinL}}) = \sum_{m,\boldsymbol{y}} P_M(m) W^n(\boldsymbol{y}|\boldsymbol{x}_m) \mathbf{1}\{m \neq \hat{m}(\boldsymbol{y})\} \tag{27}$$

$$= \underline{P_{\mathrm{e}}}(\mathcal{C}), \tag{28}$$

where $\mathbf{1}\{\cdot\}$ indicates the indicator function. As T_{MinL} is equivalent to an anti-Neyman-Pearson test, it holds that

$$\overline{\beta}_{\underline{P_{\mathrm{e}}}(\mathcal{C})}(P_M W^n, P_M \underline{R}^*_{\boldsymbol{Y}}) = \frac{1}{M_n}, \tag{29}$$

$$\underline{P_{\mathrm{e}}}(\mathcal{C}) = \overline{\alpha}_{\frac{1}{M_n}}(P_M W^n, P_M \underline{R}^*_{\boldsymbol{Y}}), \tag{30}$$

where all the quantities $\underline{P_{\mathrm{e}}}(\mathcal{C})$, W^n, $\underline{R}^*_{\boldsymbol{Y}}$, and M_n depend on the code $\mathcal{C}$. Besides, the error probability $\underline{P_{\mathrm{e}}}(\mathcal{C})$ under MinL decoding lies in the interval $1 - \frac{1}{M_n} \leq \underline{P_{\mathrm{e}}}(\mathcal{C}) \leq 1$.

To derive general bounds on the unreliability function, we now consider mismatched tests of the form T_{MinL} with alternative distributions $R_{\boldsymbol{Y}}$, different from $\underline{R}^*_{\boldsymbol{Y}}$ in general and possibly unrelated to the code. If $m = \hat{m}(\boldsymbol{y})$ the test outputs $P_M W^n$ and $P_M R_{\boldsymbol{Y}}$ otherwise. Now, Eqs. (26)–(28) remain unchanged but Eqs. (29)–(30) are replaced by

$$\overline{\beta}_{\underline{P_{\mathrm{e}}}(\mathcal{C})}(P_M W^n, P_M R_{\boldsymbol{Y}}) \geq \frac{1}{M_n}, \tag{31}$$

$$\underline{P_{\mathrm{e}}}(\mathcal{C}) \leq \overline{\alpha}_{\frac{1}{M_n}}(P_M W^n, P_M R_{\boldsymbol{Y}}). \tag{32}$$

These equations are reminiscent of the meta-converse equations for MaxL decoding, that is,

$$\underline{\beta}_{\overline{P_{\mathrm{e}}}(\mathcal{C})}(P_M W^n, P_M R_{\boldsymbol{Y}}) \leq \frac{1}{M_n}, \tag{33}$$

$$\overline{P_{\mathrm{e}}}(\mathcal{C}) \geq \underline{\alpha}_{\frac{1}{M_n}}(P_M W^n, P_M R_{\boldsymbol{Y}}). \tag{34}$$

Differently from the metaconverse for MaxL decoding, where the error probability $\overline{P_{\mathrm{e}}}(\mathcal{C}) \leq 1 - \frac{1}{M_n}$, the metaconverse for MinL decoding is valid for $\overline{P_{\mathrm{e}}}(\mathcal{C}) \geq 1 - \frac{1}{M_n}$.

Using that $\underline{\beta}_\alpha(A, B) + \overline{\beta}_{1-\alpha}(A, B) = 1$ and its equivalent for $\underline{\alpha}$ and $\overline{\alpha}$, Eqs. (31) and (32) can be rewritten in terms of the correct-decoding probability $\underline{P_{\mathrm{c}}}(\mathcal{C})$ as

$$\underline{\beta}_{\underline{P_{\mathrm{c}}}(\mathcal{C})}(P_M W^n, P_M R_{\boldsymbol{Y}}) \leq 1 - \frac{1}{M_n}, \tag{35}$$

$$\underline{P_{\mathrm{c}}}(\mathcal{C}) \geq \underline{\alpha}_{1-\frac{1}{M_n}}(P_M W^n, P_M R_{\boldsymbol{Y}}). \tag{36}$$

These metaconverse bounds establish fundamental limits on the correct-decoding probability under minimum likelihood decoding and codes with M_n codewords. Mirroring the approach used for maximum likelihood, we derive an upper bound on the unreliability function $\underline{E}(R)$ in the following section by exploiting the identity between Gallager's E_0-function and a Rényi divergence of order $\frac{1}{1+\rho}$ and applying the data processing inequality to the divergence.

IV. Product distributions as extremes of the E_0 function

In addition to the identity in Eq. (11), which connects the function $E_0(\rho, Q)$ and the Rényi divergence of order $\frac{1}{1+\rho}$, Sibson introduced in [5] the information measure $K_{\frac{1}{1+\rho}}(Q, W)$, which is defined as

$$K_{\frac{1}{1+\rho}}(Q, W) = D_{\frac{1}{1+\rho}}\big(QW \| QR_Y^{\rho,Q}\big). \tag{37}$$

Now, for an arbitrary distribution R_Y, using the identity

$$D_{\frac{1}{1+\rho}}(QW \| QR_Y) = D_{\frac{1}{1+\rho}}\big(QW \| QR_Y^{\rho,Q}\big) + D_{\frac{1}{1+\rho}}\big(QR_Y^{\rho,Q} \| QR_Y\big), \tag{38}$$

together with the fact that the Rényi divergence of order $\frac{1}{1+\rho}$ is non-negative for $\rho > -1$ and non-positive for $\rho < -1$, Sibson's measure $K_{\frac{1}{1+\rho}}(Q, W)$ can be written as the solution [6] of an optimization problem:

$$K_{\frac{1}{1+\rho}}(Q, W) = \begin{cases} \inf_{R_Y} D_{\frac{1}{1+\rho}}(QW \| QR_Y) & \rho > -1, \\ \sup_{R_Y} D_{\frac{1}{1+\rho}}(QW \| QR_Y) & \rho < -1. \end{cases} \tag{39}$$

For $\rho > -1$, Csiszár proved in [6] that the optimization over input distributions of the information measure $K_{\frac{1}{1+\rho}}(Q, W)$ admits an equivalent characterization as a double optimization over output distributions R_Y and input symbols x, namely

$$\sup_Q K_{\frac{1}{1+\rho}}(Q, W) = \inf_{R_Y} \sup_x D_{\frac{1}{1+\rho}}(W_x \| R_Y). \tag{40}$$

Using this equivalence, Polyanskiy and Verdú proved that the extremal values of Gallager's E_0-function for product channels are product distributions for $\rho > -1$. Specifically, let Q^n, W^n, and $R_{\mathbf{Y}}$ be the length-n input, conditional channel, and output distributions, respectively. Then, [4, Eqs. (45)–(48)] imply that

$$\sup_{Q^n} K_{\frac{1}{1+\rho}}(Q^n, W^n) = n \sup_Q K_{\frac{1}{1+\rho}}(Q, W), \tag{41}$$

and also that the optimum value in the maximization of $K_{\frac{1}{1+\rho}}(Q^n, W^n)$ over Q^n is attained by product distributions.

For $\rho < -1$, as explained in detail in Appendix A, Csiszár's and Polyanskiy and Verdú's proofs can be adapted to yield

$$\inf_{Q^n} K_{\frac{1}{1+\rho}}(Q^n, W^n) = n \inf_Q K_{\frac{1}{1+\rho}}(Q, W). \tag{42}$$

$$\sup_{Q^n} E_0(\rho, Q^n) = n \sup_Q E_0(\rho, Q). \tag{43}$$

V. Upper bound to the unreliability function

Let $P_{\mathrm{c}} = \underline{P_{\mathrm{c}}}(\mathcal{C})$ and $d_{\frac{1}{1+\rho}}(A \| B)$ denote the binary Rényi divergence of order $\frac{1}{1+\rho}$ between A and B. Then, the data-processing inequality for the Rényi divergence for the region $\rho < -1$ gives

$$D_{\frac{1}{1+\rho}}(P_M W^n \,\|\, P_M R_{\mathbf{Y}}^{\rho,P_M}) \le d_{\frac{1}{1+\rho}}\big(P_{\mathrm{c}} \,\|\, \overline{\beta}_{\underline{P_{\mathrm{c}}}(\mathcal{C})}(P_M W^n, P_M R_{\mathbf{Y}}^{\rho,P_M})\big) \tag{44}$$

$$\le d_{\frac{1}{1+\rho}}\big(P_{\mathrm{c}} \,\|\, \tfrac{1}{M_n}\big), \tag{45}$$

where we used (31) to obtain the last inequality. Using the definition of the binary Rényi divergence (see, e.g., [4, Eq. (24)]), we obtain that

$$d_{\frac{1}{1+\rho}}\Big(P_{\mathrm{c}} \,\|\, \tfrac{1}{M_n}\Big) = \log M_n - \frac{1}{\rho}\log P_{\mathrm{c}} - \frac{1+\rho}{\rho}\log\left(1 + \left(\frac{1-P_{\mathrm{c}}}{P_{\mathrm{c}}}\right)^{\frac{1}{1+\rho}} (M_n - 1)^{\frac{\rho}{1+\rho}}\right). \tag{46}$$

Now, in the regions $0 \le P_{\mathrm{c}} \le \frac{1}{M_n}$, $M_n \ge 2$, we have

$$\sup_{P_{\mathrm{c}}, M_n} \left\{ -\frac{1+\rho}{\rho}\log\left(1 + \left(\frac{1-P_{\mathrm{c}}}{P_{\mathrm{c}}}\right)^{\frac{1}{1+\rho}} (M_n - 1)^{\frac{\rho}{1+\rho}}\right)\right\} = 0, \tag{47}$$

which implies that

$$D_{\frac{1}{1+\rho}}(P_M W^n \,\|\, P_M R_{\mathbf{Y}}^{\rho,P_M}) \le \log M_n - \frac{1}{\rho}\log P_{\mathrm{c}}. \tag{48}$$

Finally, using the relationship between Rényi divergence and the E_0 function, we find that the correct-decoding probability $P_{\mathrm{c}} = \underline{P_{\mathrm{c}}}(\mathcal{C})$ for the code $\mathcal{C}$ with distribution P_M over a given set of codewords $\boldsymbol{x}_m$ satisfies

$$\underline{P_{\mathrm{c}}}(\mathcal{C}) \ge M_n^{\rho} F_0(\rho, P_M) \tag{49}$$

$$= e^{-(E_0(\rho, P_M) - n\rho R)}. \tag{50}$$

Since the bound is valid for any $\rho < -1$ and code distribution P_M, we have that the lowest possible correct-decoding probability for given n and M_n, $\underline{P_{\mathrm{c}}^*}(n, M_n)$, satisfies

$$\underline{P_{\mathrm{c}}^*}(n, M_n) \ge \min_{P_M} \sup_{\rho < -1} e^{-(E_0(\rho, P_M) - n\rho R)}. \tag{51}$$

The bound can be relaxed further by considering the minimum over distributions over the (larger) set of codewords $Q_{\boldsymbol{X}}$ instead of message (or code) distributions P_M, that is

$$\underline{P_{\mathrm{c}}^*}(n, M_n) \ge \min_{Q_{\boldsymbol{X}}} \sup_{\rho < -1} e^{-(E_0(\rho, Q_{\boldsymbol{X}}) - n\rho R)}. \tag{52}$$

Next, interchanging the order of the optimizations further weakens the bound, that is

$$\underline{P_{\mathrm{c}}^*}(n, M_n) \ge \sup_{\rho < -1} \min_{Q_{\boldsymbol{X}}} e^{-(E_0(\rho, Q_{\boldsymbol{X}}) - n\rho R)}. \tag{53}$$

In order to obtain the exponential decay of $\underline{P_{\mathrm{c}}^*}(n, M_n)$, the lowest correct-decoding probability, we take logarithms on both sides, divide by n, and use the fact that the extreme points for Gallager's E_0 are achieved by product distributions, i.e., Eq. (43) proved in the Appendix. Therefore, the channel unreliability function $\underline{E}(R)$, the exponential decay of P_{c} for any code of rate R under minimum likelihood decoding is upper bounded in terms of Gallager's $E_0(\rho)$ function in the region $\rho < -1$ by

$$\underline{E}(R) \le \inf_{\rho < -1} \max_Q \big\{ E_0(\rho, Q) - \rho R \big\}. \tag{54}$$

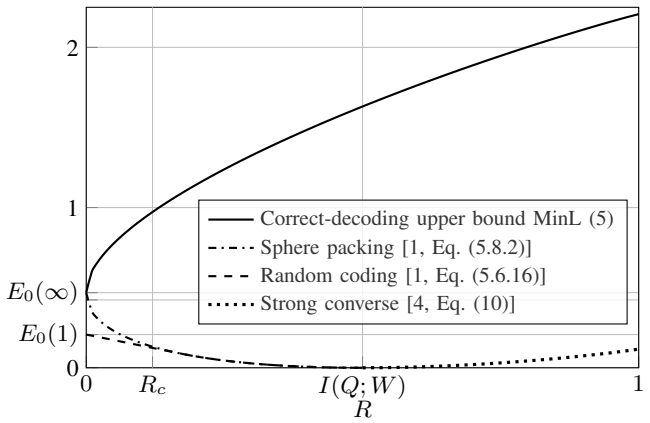

Fig. 1. Exponents vs. rate R for a BSC with crossover $p = 0.11$ and input probability $q = 0.5$. The capacity is $C = 0.500084$, the cutoff rate is $E_0(1) = 0.20716$, the critical rate is $R_c = E_0'(1) = 0.11997$, and the $\pm$-infinite-ρ plateau is $E_0(\infty) = 0.468757$.

VI. Numerical Example and Final Remarks

The formula in (5) is reminiscent of the standard MaxL random-coding [1, Eq. (5.6.16)], sphere-packing [1, Eq. (5.8.2)], and strong-converse exponents [4, Eq. (10)] at rate R, each optimizing a function of a Rényi divergence of positive order. The exponent to the converse bound derived in this paper gives operational meaning to Gallager's $E_0(\rho, Q)$ function in the region $\rho < -1$ and to Rényi divergences of negative order. We conclude with some numerical results for a binary symmetric channel: Fig. 1 shows these upper and lower bounds under MaxL decoding, together with the upper bound to the exponential decay of the worst correct-decoding probability MinL (5), and depicts a smooth transition between the sphere-packing and the MinL exponents at zero rate.

Appendix

Let $\gamma_\rho = -\frac{1+\rho}{\rho}$. First, using the definition of the Rényi divergence in (39) for $\rho < -1$ gives the following expressions for $K_{\frac{1}{1+\rho}}(Q, W)$,

$$\gamma_\rho \log\left(-\sup_{R_Y}\left\{-\sum_{x,y} Q(x)W(y|x)^{\frac{1}{1+\rho}} R_Y(y)^{\frac{\rho}{1+\rho}}\right\}\right), \quad (55)$$

where we have moved the optimization inside the logarithm, taking care of the effect of the sign of γ_ρ. Taking now the minimization over Q, $\inf_Q K_{\frac{1}{1+\rho}}(Q, W)$, moving it inside the logarithm, and applying the minimax theorem, as the objective function is linear in Q and concave in R_Y, yields

$$\begin{aligned}&\gamma_\rho \log\left(-\sup_{R_Y}\inf_{Q}\left\{-\sum_{x,y} Q(x)W(y|x)^{\frac{1}{1+\rho}} R_Y(y)^{\frac{\rho}{1+\rho}}\right\}\right)\\ &\quad = \gamma_\rho \log\left(-\sup_{R_Y}\inf_{x}\left\{-\sum_{y} W(y|x)^{\frac{1}{1+\rho}} R_Y(y)^{\frac{\rho}{1+\rho}}\right\}\right),\end{aligned} \quad (56)$$

where Eq. (56) is a consequence that the infimum of a linear function over the probability simplex (a convex set) is achieved by a distribution that assigns probability 1 to a single symbol x. Moving the double optimization back to the front, and using again the definition of the Rényi divergence yields

$$\begin{aligned}&\inf_Q K_{\frac{1}{1+\rho}}(Q, W)\\ &\quad = \sup_{R_Y}\inf_x\left\{\gamma_\rho \log\left(\sum_y W(y|x)^{\frac{1}{1+\rho}} R_Y(y)^{\frac{\rho}{1+\rho}}\right)\right\} \quad (57)\\ &\quad = \sup_{R_Y}\inf_x D_{\frac{1}{1+\rho}}(W_x \| R_Y), \quad (58)\end{aligned}$$

where Eq. (56) follows from the fact that the infimum of a linear function over a convex set (the probability simplex) is achieved at one of the extreme points of the set, in this case the distributions that assign probability 1 to a single symbol x, Eq. (57) moves the double optimization back to the front, and Eq. (58) is again the definition of the Rényi divergence.

Let Q^n, W^n, and $R_{\boldsymbol{Y}}$ be the length-n input, conditional channel, and output distributions, respectively. From Eq. (58), and lower bounding the optimization over $R_{\boldsymbol{Y}}$ by one over $R_{Y_1} \cdots R_{Y_n}$, we obtain the following chain of (in)equalities

$$\inf_{Q^n} K_{\frac{1}{1+\rho}}(Q^n, W^n) \geq \sup_{R_{Y_1}\cdots R_{Y_n}} \inf_{\boldsymbol{x}} D_{\frac{1}{1+\rho}}(W^n_{\boldsymbol{x}} \| R_{Y_1} \cdots R_{Y_n}) \quad (59)$$

$$= \sum_{i=1}^n \sup_{R_{Y_i}} \inf_{x_i} D_{\frac{1}{1+\rho}}(W_{x_i} \| R_{Y_i}) \quad (60)$$

$$= \sum_{i=1}^n \inf_{Q_i} K_{\frac{1}{1+\rho}}(Q_i, W) \quad (61)$$

$$= n \inf_Q K_{\frac{1}{1+\rho}}(Q, W). \quad (62)$$

In parallel, taking $Q^n(\boldsymbol{x})$ to be a product of independent distributions $Q(x_i)$ shows that $\inf_{Q^n} K_{\frac{1}{1+\rho}}(Q^n, W^n) \leq n \inf_Q K_{\frac{1}{1+\rho}}(Q, W)$. Therefore, the optimum value is attained by product distributions. Finally, using the identity between Gallager's function $E_0(\rho, Q)$ and the information measure $K_{\frac{1}{1+\rho}}(Q, W)$ in Eq. (11) shows that

$$\inf_{Q^n} \frac{1}{\rho} E_0(\rho, Q^n) = n \inf_Q \frac{1}{\rho} E_0(\rho, Q). \quad (63)$$

Since ρ is negative, we thus have

$$\sup_{Q^n} E_0(\rho, Q^n) = n \sup_Q E_0(\rho, Q), \qquad \rho < -1. \quad (64)$$

References

[1] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., 1968.

[2] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab., vol. 1, Contrib. Theory Statist.*, vol. 4. University of California Press, 1961, pp. 547–562.

[3] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[4] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, 2010, pp. 1327–1333.

[5] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969.

[6] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 26–34, 1995.