entropy

MDPI

Article

Error Exponents and α-Mutual Information

Sergio Verdú

Independent Researcher, Princeton, NJ 08540, USA; verdu@informationtheory.org

Abstract: Over the last six decades, the representation of error exponent functions for data transmission through noisy channels at rates below capacity has seen three distinct approaches: (1) Through Gallager's E_0 functions (with and without cost constraints); (2) large deviations form, in terms of conditional relative entropy and mutual information; (3) through the α-mutual information and the Augustin–Csiszár mutual information of order α derived from the Rényi divergence. While a fairly complete picture has emerged in the absence of cost constraints, there have remained gaps in the interrelationships between the three approaches in the general case of cost-constrained encoding. Furthermore, no systematic approach has been proposed to solve the attendant optimization problems by exploiting the specific structure of the information functions. This paper closes those gaps and proposes a simple method to maximize Augustin–Csiszár mutual information of order α under cost constraints by means of the maximization of the α-mutual information subject to an exponential average constraint.

Keywords: information measures; relative entropy; Rényi divergence; mutual information; α-mutual information; Augustin–Csiszár mutual information; data transmission; error exponents; large deviations

Citation: Verdú, S. Error Exponents and α-Mutual Information. *Entropy* **2021**, *23*, 199. https://doi.org/10.3390/e23020199

Received: 7 December 2020
Accepted: 28 January 2021
Published: 5 February 2021

1. Introduction

1.1. Phase 1: The MIT School

The capacity C of a stationary memoryless channel is equal to the maximal symbol-wise input–output mutual information. Not long after Shannon [1] established this result, Rice [2] observed that, when operating at any encoding rate $R < C$, there exist codes whose error probability vanishes exponentially with blocklength, with a speed of decay that decreases as R approaches C. This early observation moved the center of gravity of information theory research towards the quest for the reliability function, a term coined by Shannon [3] to refer to the maximal achievable exponential decay as a function of R. The MIT information theory school, and most notably, Elias [4], Feinstein [5], Shannon [3,6], Fano [7], Gallager [8,9], and Shannon, Gallager and Berlekamp [10,11], succeeded in upper/lower bounding the reliability function by the sphere-packing error exponent function and the random coding error exponent function, respectively. Fortunately, these functions coincide for rates between C and a certain value, called the critical rate, thereby determining the reliability function in that region. The influential 1968 textbook by Gallager [9] set down the major error exponent results obtained during Phase 1 of research on this topic, including the expurgation technique to improve upon the random coding error exponent lower bound. Two aspects of those early works (and of Dobrushin's contemporary papers [12,13] on the topic) stand out:

(a) The error exponent functions were expressed as the result of the Karush-Kuhn-Tucker optimization of ad-hoc functions which, unlike mutual information, carried little insight. In particular, during the first phase, center stage is occupied by the parametrized function of the input distribution P_X and the random transformation (or "channel") $P_{Y|X}$,

$$E_0(\rho, P_X) = -\log \sum_{y\in\mathcal{B}} \left(\sum_{x\in\mathcal{A}} P_X(x) P_{Y|X}^{\frac{1}{1+\rho}}(y|x) \right)^{1+\rho}, \tag{1}$$

introduced by Gallager in [8].

(b) Despite the large-deviations nature of the setup, none of the tools from that then-nascent field (other than the Chernoff bound) found their way to the first phase of the work on error exponents; in particular, relative entropy, introduced by Kullback and Leibler [14], failed to put in an appearance.

To this date, the reliability function remains open for low rates even for the binary symmetric channel, despite a number of refined converse and achievability results (e.g., [15–21]) obtained since [9]. Our focus in this paper is not on converse/achievability techniques but on the role played by various information measures in the formulation of error exponent results.

1.2. Phase 2: Relative Entropy

The second phase of the error exponent research was pioneered by Haroutunian [22] and Blahut [23], who infused the expressions for the error exponent functions with meaning by incorporating relative entropy. The sphere-packing error exponent function corresponding to a random transformation $P_{Y|X}$ is given as

$$E_{\mathtt{sp}}(R) = \sup_{P_X} \min_{\substack{Q_{Y|X}\colon \mathcal{A}\to\mathcal{B} \\ I(P_X, Q_{Y|X}) \leqslant R}} D(Q_{Y|X}\|P_{Y|X}|P_X). \tag{2}$$

Roughly speaking, optimal codes of rate $R < C$ incur in errors due to atypical channel behavior, and large deviations establishes that the overwhelmingly most likely such behavior can be explained as if the channel would be supplanted by the one with mutual information bounded by R which is closest to the true channel in conditional relative entropy $D(Q_{Y|X}\|P_{Y|X}|P_X)$. Within the confines of finite-alphabet memoryless channels, this direction opened the possibility of using the combinatorial method of types to obtain refined results robustifying the choice of the optimal code against incomplete knowledge of the channel. The 1981 textbook by Csiszár and Körner [24] summarizes the main results obtained during Phase 2.

1.3. Phase 3: Rényi Information Measures

Entropy and relative entropy were generalized by Rényi [25], who introduced the notions of Rényi entropy and Rényi divergence of order α. He arrived at Rényi entropy by relaxing the axioms Shannon proposed in [1], and showed to be satisfied by no measure but entropy. Shortly after [25], Campbell [26] realized the operational role of Rényi entropy in variable-length data compression if the usual average encoding length criterion $\mathbb{E}[\ell(c(X))]$ is replaced by an exponential average $\alpha^{-1}\log\mathbb{E}[\exp(\alpha\,\ell(c(X))]$. Arimoto [27] put forward a generalized conditional entropy inspired by Rényi's measures (now known as Arimoto-Rényi conditional entropy) and proposed a generalized mutual information by taking the difference between Rényi entropy and the Arimoto-Rényi conditional entropy. The role of the Arimoto-Rényi conditional entropy in the analysis of the error probability of Bayesian M-ary hypothesis testing problems has been recently shown in [28], tightening and generalizing a number of results dating back to Fano's inequality [29].

Phase 3 of the error exponent research was pioneered by Csiszár [30] where he established a connection between Gallager's E_0 function and Rényi divergence by means of a Bayesian measure of the discrepancy among a finite collection of distributions introduced by Sibson [31]. Although [31] failed to realize its connection to mutual information, Csiszár [30,32] noticed that it could be viewed as a natural generalization of mutual information. Arimoto [27] also observed that the unconstrained maximization of his generalized mutual information measure with respect to the input distribution coincides with a scaled version of the maximal E_0 function. This resulted in an extension of the Arimoto-Blahut algorithm useful for the computation of error exponent functions [33] (see also [34]) for finite-alphabet memoryless channels.

Within Haroutunian's framework [22] applied in the context of the method of types, Poltyrev [35] proposed an alternative to Gallager's E_0 function, defined by means of a cumbersome maximization over a reverse random transformation. This measure turned out to coincide (modulo different parametrizations) with another generalized mutual information introduced four years earlier by Augustin in his unpublished thesis [36], by means of a minimization with respect to an output probability measure.

The key contribution in the development of this third phase is Csiszár's paper [32] where he makes a compelling case for the adoption of Rényi's information measures in the large deviations analysis of lossless data compression, hypothesis testing and data transmission. Recall that more than two decades earlier, Csiszár [30] had already established the connection of Gallager's E_0 function and the generalized mutual information inspired by Sibson [31], which, henceforth, we refer to as the α-mutual information. Therefore, its relevance to the error exponent analysis of error correcting codes had already been established. Incidentally, more recently, another operational role was found for α-mutual information in the context of the large deviations analysis of composite hypothesis testing [37]. In addition to α-mutual information, and always working with discrete alphabets, Csiszár [32] considers the generalized mutual informations due to Arimoto [27], and to Augustin [36], which we refer to as the Augustin–Csiszár mutual information of order α. Csiszár shows that all those three generalizations of mutual information coincide upon their unconstrained maximization with respect to the input distribution. Further relationships among those Rényi-based generalized mutual informations have been obtained in recent years in [38–45]. In [32] the maximal α-mutual information or generalized capacity of order α finds an operational characterization as a generalized cutoff rate–an equivalent way to express the reliability function. This would have been the final word on the topic if it weren't for its limitation to discrete-alphabet channels, and more importantly, encoding without cost constraints.

1.4. Cost Constraints

If the transmitted codebook is cost-constrained, i.e., every codeword $(c_1, \ldots, c_n)$ is forced to satisfy $\sum_{i=1}^{n} \mathsf{b}(c_i) \leqslant n\,\theta$ for some nonnegative cost function $\mathsf{b}(\cdot)$, then the channel capacity is equal to the input–output mutual information maximized over input probability measures restricted to satisfy $\mathbb{E}[\mathsf{b}(X)] \leqslant \theta$. Gallager [9] incorporated cost constraints in his treatment of error exponents by generalizing (1) to the function

$$E_0(\rho, P_X, r, \theta) = -\log \sum_{y \in \mathcal{B}} \left(\sum_{x \in \mathcal{A}} P_X(x) \exp(r\,\mathsf{b}(x) - r\,\theta) P_{Y|X}^{\frac{1}{1+\rho}}(y|x) \right)^{1+\rho}, \tag{3}$$

with which he was able to prove an achievability result invoking Shannon's random coding technique [1]. Gallager also suggested in the footnote of page 329 of [9] that the converse technique of [10] is amenable to extension to prove a sphere-packing converse based on (3). However, an important limitation is that that technique only applies to constant-composition codes (all codewords have the same empirical distribution). A more powerful converse circumventing that limitation (at least for symmetric channels) was given by [46] also expressing the upper bound on the reliability function by optimizing (3) with respect

to ρ, r and P_X. A notable success of the approach based on the optimization of (3) was the determination of the reliability function (for all rates below capacity) of the direct detection photon channel [47].

In contrast, the Phase Two expression (2) for the sphere-packing error exponent for cost-constrained channels is much more natural and similar to the way the expression for channel capacity is impacted by cost constraints, namely we simply constrain the maximization in (2) to satisfy $\mathbb{E}[\mathsf{b}(X)] \leqslant \theta$. Unfortunately, no general methods to solve the ensuing optimization have been reported.

Once cost constraints are incorporated, the equivalence among the maximal α-mutual information, maximal order-α Augustin–Csiszár mutual information, and maximal Arimoto mutual information of order α breaks down. Of those three alternatives, it is the maximal Augustin–Csiszár mutual information under cost constraints that appears in the error exponent functions. The challenge is that Augustin–Csiszár mutual information is much harder to evaluate, let alone maximize, than α-mutual information. The Phase 3 effort to encompass cost constraints started by Augustin [36] and was continued recently by Nakiboglu [43]. Their focus was to find a way to express (3) in terms of Rényi information measures. Although, as we explain in Item 62, they did not quite succeed, their efforts were instrumental in developing key properties of the Augustin–Csiszár mutual information.

1.5. Organization

To enhance readability and ease of reference, the rest of this work is organized in 81 items, grouped into Section 13 and an appendix.

Basic notions and notation (including the key concept of α-response) are collected in Section 2. Unlike much of the literature on the topic, we do not restrict attention to discrete input/output alphabets, nor do we impose any topological structures on them.

The paper is essentially self-contained. Section 3 covers the required background material on relative entropy, Rényi divergence of order α, and their conditional versions, including a key representation of Rényi divergence in terms of relative entropies and a tilted probability measure, and additive decompositions of Rényi divergence involving the α-response.

Section 4 studies the basic properties of α-mutual information and order-α Augustin–Csiszár mutual information. This includes their variational representations in terms of conventional (non-Rényi) information measures such as conditional relative entropy and mutual information, which are particularly simple to show in the main range of interest in applications to error exponents, namely, $\alpha \in (0,1)$.

The interrelationships between α-mutual information and order-α Augustin–Csiszár mutual information are covered in Section 5, which introduces the dual notions of α-adjunct and $\langle\alpha\rangle$-adjunct of an input probability measure.

The maximizations with respect to the input distribution of α-mutual information and order-α Augustin–Csiszár mutual information account for their role in the fundamental limits in data transmission through noisy channels. Section 6 gives a brief review of the results in [45] for the maximization of α-mutual information. For Augustin–Csiszár mutual information, Section 7 covers its unconstrained maximization, which coincides with its α-mutual information counterpart. Section 8 proposes an approach to find $\mathbb{C}_\alpha^{\mathsf{c}}(\theta)$, the maximal Augustin–Csiszár mutual information of order $\alpha \in (0,1)$ subject to $\mathbb{E}[\mathsf{b}(X)] \leqslant \theta$. Instead of trying to identify directly the input distribution that maximizes Augustin–Csiszár mutual information, the method seeks its $\langle\alpha\rangle$-adjunct. This is tantamount to maximizing α-mutual information over a larger set of distributions.

Section 9 shows

$$\rho\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\theta) = \min_{r\geqslant 0} \max_{P_X} E_0(\rho, P_X, r, \theta), \tag{4}$$

where the maximization on the right side is unconstrained. In other words, the minimax of Gallager's E_0 function (3) with cost constraints is shown to be equal to the maximal

Augustin–Csiszár mutual information, thereby bridging the existing gap between the Phase 1 and Phase 3 representations alluded to earlier in this introduction.

As in [48], Section 10 defines the sphere-packing and random-coding error exponent functions in the natural canonical form of Phase 2 (e.g., (2)), and gives a very simple proof of the nexus between the Phase 2 and Phase 3 representations, namely,

$$E_{\mathsf{sp}}(R) = \sup_{\rho \geqslant 0} \left\{ \rho \, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\theta) - \rho \, R \right\}, \tag{5}$$

with or without cost constraints. In this regard, we note that, although all the ingredients required were already present at the time the revised version of [24] was published three decades after the original, [48] does not cover the role of Rényi's information measures in channel error exponents.

Examples illustrating the proposed method are given in Sections 11 and 12 for the additive Gaussian noise channel under a quadratic cost function, and the additive exponential noise channel under a linear cost function, respectively. Simple parametric expressions are given for the error exponent functions, and the least favorable channels that account for the most likely error mechanism (Section 1.2) are identified in both cases.

2. Relative Information and Information Density

We begin with basic terminology and notation required for the subsequent development.

1. If $(\mathcal{A}, \mathscr{F}, P)$ is a probability space, $X \sim P$ indicates $\mathbb{P}[X \in \mathcal{F}] = P(\mathcal{F})$ for all $\mathcal{F} \in \mathscr{F}$.
2. If probability measures P and Q defined on the same measurable space $(\mathcal{A}, \mathscr{F})$ satisfy $P(A) = 0$ for all $A \in \mathscr{F}$ such that $Q(A) = 0$, we say that P is dominated by Q, denoted as $P \ll Q$. If P and Q dominate each other, we write $P \ll\gg Q$. If there is an event such that $P(A) = 0$ and $Q(A) = 1$, we say that P and Q are mutually singular, and we write $P \perp Q$.
3. If $P \ll Q$, then $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative of the dominated measure P with respect to the reference measure Q. Its logarithm is known as the relative information, namely, the random variable

$$\imath_{P\|Q}(a) = \log \frac{\mathrm{d}P}{\mathrm{d}Q}(a) \in [-\infty, +\infty), \quad a \in \mathcal{A}. \tag{6}$$

As with the Radon-Nikodym derivative, any identity involving relative informations can be changed on a set of measure zero under the reference measure without incurring in any contradiction. If $P \ll Q \ll R$, then the chain rule of Radon-Nikodym derivatives yields

$$\imath_{P\|Q}(a) + \imath_{Q\|R}(a) = \imath_{P\|R}(a), \quad a \in \mathcal{A}. \tag{7}$$

Throughout the paper, the base of exp and log is the same and chosen by the reader unless explicitly indicated otherwise. We frequently define a probability measure P from the specification of $\imath_{P\|Q}$ and $Q \gg P$ since

$$P(A) = \int_A \exp\Big(\imath_{P\|Q}(a)\Big) \, \mathrm{d}Q(a), \quad A \in \mathscr{F}. \tag{8}$$

If $X \sim P$ and $Y \sim Q$, it is often convenient to write $\imath_{X\|Y}(x)$ instead of $\imath_{P\|Q}(x)$. Note that

$$\mathbb{E}\Big[\exp\Big(\imath_{X\|Y}(Y)\Big)\Big] = 1. \tag{9}$$

Example 1. *If* $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *(Gaussian with mean* μ_X *and variance* σ_X^2*) and* $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$*, then,*

$$\imath_{X\|Y}(a) = \frac{1}{2}\log\frac{\sigma_Y^2}{\sigma_X^2} + \frac{1}{2}\left(\frac{(a-\mu_Y)^2}{\sigma_Y^2} - \frac{(a-\mu_X)^2}{\sigma_X^2}\right)\log \mathrm{e}. \tag{10}$$

4. Let $(\mathcal{A}, \mathscr{F})$ and $(\mathcal{B}, \mathscr{G})$ be measurable spaces, known as the input and output spaces, respectively. Likewise, $\mathcal{A}$ and $\mathcal{B}$ are referred to as the input and output alphabets respectively. The simplified notation $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$ denotes a random transformation from $(\mathcal{A}, \mathscr{F})$ to $(\mathcal{B}, \mathscr{G})$, i.e. for any $x \in \mathcal{A}$, $P_{Y|X=x}(\cdot)$ is a probability measure on $(\mathcal{B}, \mathscr{G})$, and for any $B \in \mathscr{G}$, $P_{Y|X=\cdot}(B)$ is an $\mathscr{F}$-measurable function.
5. We abbreviate by $\mathcal{P}_{\mathcal{A}}$ the set of probability measures on $(\mathcal{A}, \mathscr{F})$, and by $\mathcal{P}_{\mathcal{A}\times\mathcal{B}}$ the set of probability measures on $(\mathcal{A}\times\mathcal{B}, \mathscr{F}\otimes\mathscr{G})$. If $P \in \mathcal{P}_{\mathcal{A}}$ and $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$ is a random transformation, the corresponding joint probability measure is denoted by $P\,P_{Y|X} \in \mathcal{P}_{\mathcal{A}\times\mathcal{B}}$ (or, interchangeably, $P_{Y|X}P$). The notation $P \to P_{Y|X} \to Q$ simply indicates that the output marginal of the joint probability measure $P\,P_{Y|X}$ is denoted by $Q \in \mathcal{P}_{\mathcal{B}}$, namely,

$$Q(B) = \int P_{Y|X}(B|x)\,\mathrm{d}P_X(x) = \mathbb{E}\Big[P_{Y|X}(B|X)\Big], \quad B \in \mathscr{G}. \tag{11}$$

6. If $P_X \to P_{Y|X} \to P_Y$ and $P_{Y|X=a} \ll P_Y$, the information density $\imath_{X;Y}\colon \mathcal{A}\times\mathcal{B} \to [-\infty, \infty)$ is defined as

$$\imath_{X;Y}(a;b) = \imath_{P_{Y|X=a}\|P_Y}(b), \quad (a,b) \in \mathcal{A}\times\mathcal{B}. \tag{12}$$

Following Rényi's terminology [49], if $P_X P_{Y|X} \ll P_X \times P_Y$, the dependence between X and Y is said to be regular, and the information density can be defined on $(x,y) \in \mathcal{A}\times\mathcal{B}$. Henceforth, we assume that $P_{Y|X}$ is such that the dependence between its input and output is regular regardless of the input probability measure. For example, if $X = Y \in \mathbb{R}$, then $P_{Y|X=a}(A) = 1\{a \in A\}$, and their dependence is not regular, since for any P_X with non-discrete components $P_{XY} \not\ll P_X \times P_Y$.
7. Let $\alpha > 0$, and $P_X \to P_{Y|X} \to P_Y$. The α-response to $P_X \in \mathcal{P}_{\mathcal{A}}$ is the output probability measure $P_{Y[\alpha]} \ll P_Y$ with relative information given by

$$\imath_{Y[\alpha]\|Y}(y) = \frac{1}{\alpha}\log\mathbb{E}[\exp(\alpha\,\imath_{X;Y}(X;y) - \kappa_\alpha)], \quad X \sim P_X, \tag{13}$$

where κ_α is a scalar that guarantees that $P_{Y[\alpha]}$ is a probability measure. Invoking (9), we obtain

$$\kappa_\alpha = \alpha\log\mathbb{E}\Big[\mathbb{E}^{\frac{1}{\alpha}}[\exp(\alpha\,\imath_{X;Y}(X;\bar{Y}))|\bar{Y}]\Big], \quad (X,\bar{Y}) \sim P_X \times P_Y. \tag{14}$$

For brevity, the dependence of κ_α on P_X and $P_{Y|X}$ is omitted. Jensen's inequality applied to $(\cdot)^\alpha$ results in $\kappa_\alpha \leqslant 0$ for $\alpha \in (0,1)$ and $\kappa_\alpha \geqslant 0$ for $\alpha > 1$. Although the α-response has a long record of services to information theory, this terminology and notation were introduced recently in [45]. Alternative terminology and notation were proposed in [42], which refers to the α-response as the order α Rényi mean. Note that $\kappa_1 = 0$ and the 1-response to P_X is P_Y. If $p_{Y[\alpha]}$ and $p_{Y|X}$ denote the densities of $P_{Y[\alpha]}$ and $P_{Y|X}$ with respect to some common dominating measure, then (13) becomes

$$p_{Y[\alpha]}(y) = \exp\Big(-\frac{\kappa_\alpha}{\alpha}\Big)\,\mathbb{E}^{\frac{1}{\alpha}}\Big[p_{Y|X}^\alpha(y|X)\Big], \quad X \sim P_X. \tag{15}$$

For $\alpha > 1$ (resp. $\alpha < 1$) we can think of the normalized version of $p_{Y|X}^\alpha$ as a random transformation with less (resp. more) "noise" than $p_{Y|X}$.

8. We will have opportunity to apply the following examples.

Example 2. *If $Y = X + N$, where $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ independent of $N \sim \mathcal{N}(\mu_N, \sigma_N^2)$, then the α-response to P_X is*

$$Y[\alpha] \sim \mathcal{N}\Big(\mu_X + \mu_N, \alpha\,\sigma_X^2 + \sigma_N^2\Big). \tag{16}$$

Example 3. *Suppose that $Y = X + N$, where N is exponential with mean ζ, independent of X, which is a mixed random variable with density*

$$f_X(t) = \frac{\zeta}{\alpha\,\mu}\delta(t) + \left(1 - \frac{\zeta}{\alpha\,\mu}\right)\frac{1}{\mu}\mathrm{e}^{-t/\mu}\,1\{t > 0\}, \tag{17}$$

with $\alpha\,\mu \geqslant \zeta$. Then, $Y[\alpha]$, the α-response to P_X, is exponential with mean $\alpha\,\mu$.

3. Relative Entropy and Rényi Divergence

Given a pair of probability measures $(P, Q) \in \mathcal{P}_{\mathcal{A}}^2$, relative entropy and Rényi divergence gauge the distinctness between P and Q.

9. Provided $P \ll Q$, the relative entropy is the expectation of the relative information with respect to the dominated measure

$$\begin{aligned} D(P\|Q) &= \mathbb{E}\Big[\imath_{P\|Q}(X)\Big], \quad X \sim P && (18)\\ &= \mathbb{E}\Big[\exp\Big(\imath_{P\|Q}(Y)\Big)\imath_{P\|Q}(Y)\Big], \quad Y \sim Q && (19)\\ &\geqslant 0, && (20) \end{aligned}$$

with equality if and only if $P = Q$. If $P \not\ll Q$, then $D(P\|Q) = \infty$. As in Item 3, if $X \sim P$ and $Y \sim Q$, we may write $D(X\|Y)$ instead of $D(P\|Q)$, in the same spirit that the expectation and entropy of P are written as $\mathbb{E}[X]$ and $H(X)$, respectively.

10. Arising in the sequel, a common optimization in information theory finds, among the probability measures satisfying an average cost constraint, that which is closest to a given reference measure Q in the sense of $D(\cdot\|Q)$. For that purpose, the following result proves sufficient. Incidentally, we often refer to unconstrained maximizations over probability distributions. It should be understood that those optimizations are still constrained to the sets $\mathcal{P}_{\mathcal{A}}$ or $\mathcal{P}_{\mathcal{B}}$. As customary in information theory, we will abbreviate $\max_{P_X \in \mathcal{P}_{\mathcal{A}}}$ by $\max_X$ or $\max_{P_X}$.

Theorem 1. *Let $P_Z \in \mathcal{P}_{\mathcal{A}}$ and suppose that $g\colon \mathcal{A} \to [0, \infty)$ is a Borel measurable mapping. Then,*

$$\min_X\{D(X \,\|\, Z) + \mathbb{E}[g(X)]\} = -\log \mathbb{E}[\exp(-g(Z))], \tag{21}$$

achieved uniquely by $P_X^* \ll\!\gg P_Z$ defined by

$$\imath_{X^*\|Z}(a) = -g(a) - \log \mathbb{E}[\exp(-g(Z))], \quad a \in \mathcal{A}. \tag{22}$$

Proof. Note that since g is nonnegative, $\eta = \mathbb{E}[\exp(-g(Z))] \in (0, 1]$. Furthermore,

$$\mathbb{E}[g(X^*)] = \frac{\int g(t)\exp(-g(t))\,\mathrm{d}P_Z(t)}{\mathbb{E}[\exp(-g(Z))]} \in \left[0, \frac{1}{\mathrm{e}\,\eta}\right]. \tag{23}$$

Therefore, the subset of $\mathcal{P}_{\mathcal{A}}$ for which the term in $\{\cdot\}$ in (21) is finite is nonempty: Fix any P_X from that subset, (which therefore satisfies $P_X \ll P_Z \ll P_X^*$) and invoke the chain rule (7) to write

$$D(X\,\|\,Z) + \mathbb{E}[g(X)] = \mathbb{E}\Big[\imath_{X\|X^*}(X) + \imath_{X^*\|Z}(X) + g(X)\Big] \tag{24}$$

$$= D(X\,\|\,X^*) - \log \mathbb{E}[\exp(-g(Z))], \quad X \sim P_X, \tag{25}$$

which is uniquely minimized by letting $P_X = P_X^*$. Note that for typographical convenience we have denoted $X^* \sim P_X^*$. □

11. Let p and q denote the Radon-Nikodym derivatives of probability measures P and Q, respectively, with respect to a common dominating σ-finite measure μ. The Rényi divergence of order $\alpha \in (0,1) \cup (1,\infty)$ between P and Q is defined as [25,50]

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \int_{\mathcal{A}} p^\alpha q^{1-\alpha} \mathrm{d}\mu \tag{26}$$

$$= \frac{1}{\alpha - 1} \log \mathbb{E}\Big[\exp\Big(\alpha\, \imath_{P\|R}(Z) + (1-\alpha)\imath_{Q\|R}(Z)\Big)\Big], \quad Z \sim R \tag{27}$$

$$= \frac{1}{\alpha - 1} \log \mathbb{E}\Big[\exp\Big(\alpha\, \imath_{P\|Q}(Y)\Big)\Big], \quad Y \sim Q \tag{28}$$

$$= \frac{1}{\alpha - 1} \log \mathbb{E}\Big[\exp\Big((\alpha-1)\, \imath_{P\|Q}(X)\Big)\Big], \quad X \sim P, \tag{29}$$

where (28) and (29) hold if $P \not\ll Q$, and in (27), R is a probability measure that dominates both P and Q. Note that (28) and (29) state that $(t-1)D_t(X\|Y)$ and $t\, D_{1+t}(X\|Y)$ are the cumulant generating functions of the random variables $\imath_{X\|Y}(Y)$ and $\imath_{X\|Y}(X)$, respectively. The relative entropy is the limit of $D_\alpha(P\|Q)$ as $\alpha \uparrow 1$, so it is customary to let $D_1(P\|Q) = D(P\|Q)$. For any $\alpha > 0$, $D_\alpha(P\|Q) \geqslant 0$ with equality if and only if $P = Q$. Furthermore, $D_\alpha(P\|Q)$ is non-decreasing in α, satisfies the skew-symmetric property

$$(1-\alpha) D_\alpha(P\|Q) = \alpha\, D_{1-\alpha}(Q\|P), \quad \alpha \in [0,1], \tag{30}$$

and

$$\inf_{\alpha\in(0,1)} D_\alpha(P\|Q) = \infty \iff P \perp Q \Longrightarrow \inf_{\alpha>1} D_\alpha(P\|Q) = \infty. \tag{31}$$

12. The expressions in the following pair of examples will come in handy in Sections 11 and 12.

Example 4. *Suppose that* $\sigma_\alpha^2 = \alpha\,\sigma_1^2 + (1-\alpha)\sigma_0^2 > 0$ *and* $\alpha \in (0,1) \cup (1,\infty)$. *Then,*

$$D_\alpha\Big(\mathcal{N}\Big(\mu_0,\sigma_0^2\Big)\,\|\,\mathcal{N}\Big(\mu_1,\sigma_1^2\Big)\Big) = \frac{1}{2}\log\frac{\sigma_1^2}{\sigma_0^2} + \frac{1}{2(\alpha-1)}\log\frac{\sigma_1^2}{\sigma_\alpha^2} + \frac{\alpha(\mu_1-\mu_0)^2}{2\,\sigma_\alpha^2}\log \mathrm{e}, \tag{32}$$

$$D\Big(\mathcal{N}\Big(\mu_0,\sigma_0^2\Big)\,\|\,\mathcal{N}\Big(\mu_1,\sigma_1^2\Big)\Big) = \frac{1}{2}\log\frac{\sigma_1^2}{\sigma_0^2} + \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} - 1\right)\log \mathrm{e} + \frac{(\mu_1-\mu_0)^2}{2\,\sigma_1^2}\log \mathrm{e} \tag{33}$$

$$= \lim_{\alpha\to 1} D_\alpha\Big(\mathcal{N}\Big(\mu_0,\sigma_0^2\Big)\,\|\,\mathcal{N}\Big(\mu_1,\sigma_1^2\Big)\Big). \tag{34}$$

Example 5. *Suppose Z is exponentially distributed with unit mean, i.e., its probability density function is $\mathrm{e}^{-t}1\{t \geqslant 0\}$. For $d_0 \geqslant d_1$ and α such that $(1-\alpha)\,\mu_0 + \alpha\,\mu_1 > 0$ we obtain*

$$D_\alpha(\mu_0\,Z + d_0 \,\|\, \mu_1\,Z + d_1) = \frac{d_0 - d_1}{\mu_1}\log \mathrm{e} + \log\frac{\mu_1}{\mu_0} + \frac{1}{1-\alpha}\log\left(\alpha + (1-\alpha)\frac{\mu_0}{\mu_1}\right),$$

$$D(\mu_0\,Z + d_0 \,\|\, \mu_1\,Z + d_1) = \left(\frac{\mu_0}{\mu_1} - 1 + \frac{d_0 - d_1}{\mu_1}\right)\log \mathrm{e} + \log\frac{\mu_1}{\mu_0} \tag{35}$$

$$= \lim_{\alpha\to 1} D_\alpha(\mu_0\,Z + d_0 \,\|\, \mu_1\,Z + d_1). \tag{36}$$

13. Intimately connected with the notion of Rényi divergence is the tilted probability measure P_α defined, if $D_\alpha(P_1\|P_0) < \infty$, by

$$\imath_{P_\alpha\|Q}(a) = \alpha\,\imath_{P_1\|Q}(a) + (1-\alpha)\,\imath_{P_0\|Q}(a) + (1-\alpha)D_\alpha(P_1 \,\|\, P_0), \tag{37}$$

where Q is any probability measure that dominates both P_0 and P_1. Although (37) is defined in general, our main emphasis is on the range $\alpha \in (0,1)$, in which, as long as $P_0 \not\perp P_1$, the tilted probability measure is defined and satisfies $P_\alpha \ll P_0$ and $P_\alpha \ll P_1$, with corresponding relative informations

$$\imath_{P_\alpha\|P_0}(a) = \imath_{P_\alpha\|Q}(a) - \imath_{P_0\|Q}(a) \tag{38}$$

$$= (1-\alpha)D_\alpha(P_1 \,\|\, P_0) + \alpha\left(\imath_{P_1\|Q}(a) - \imath_{P_0\|Q}(a)\right), \tag{39}$$

$$\imath_{P_\alpha\|P_1}(a) = \imath_{P_\alpha\|Q}(a) - \imath_{P_1\|Q}(a) \tag{40}$$

$$= (1-\alpha)D_\alpha(P_1 \,\|\, P_0) - (1-\alpha)\left(\imath_{P_1\|Q}(a) - \imath_{P_0\|Q}(a)\right), \tag{41}$$

where we have used the chain rule for $P_\alpha \ll P_0 \ll Q$ and $P_\alpha \ll P_1 \ll Q$. Taking a linear combination of (38)–(41) we conclude that, for all $a \in \mathcal{A}$,

$$(1-\alpha)D_\alpha(P_1\|P_0) = (1-\alpha)\,\imath_{P_\alpha\|P_0}(a) + \alpha\,\imath_{P_\alpha\|P_1}(a). \tag{42}$$

Henceforth, we focus particular attention on the case $\alpha \in (0,1)$ since that is the region of interest in the application of Rényi information measures to the evaluation of error exponents in channel coding for codes whose rate is below capacity. In addition, often proofs simplify considerably for $\alpha \in (0,1)$.

14. Much of the interplay between relative entropy and Rényi divergence hinges on the following identity, which appears, without proof, in (3) of [51].

Theorem 2. *Let $\alpha \in (0,1)$ and assume that $P_0 \not\perp P_1$ are defined on the same measurable space. Then, for any $P \ll P_1$ and $P \ll P_0$,*

$$\alpha\,D(P\,\|P_1) + (1-\alpha)\,D(P\,\|P_0) = D(P\,\|P_\alpha) + (1-\alpha)D_\alpha(P_1\,\|P_0), \tag{43}$$

where P_α is the tilted probability measure in (37) and (43) holds regardless of whether the relative entropies are finite. In particular,

$$D(P\,\|\,P_\alpha) < \infty \iff \max\{D(P\,\|\,P_0), D(P\,\|\,P_1)\} < \infty. \tag{44}$$

Proof. We distinguish three overlapping cases:

(1) $D(P\,\|\,P_\alpha) < \infty$: Taking expectation of (42) with respect to $a \leftarrow X \sim P$, yields (43) because

$$\mathbb{E}\Big[\imath_{P_\alpha\|P_0}(X)\Big] = D(P\|P_0) - D(P\|P_\alpha), \tag{45}$$

$$\mathbb{E}\Big[\imath_{P_\alpha\|P_1}(X)\Big] = D(P\|P_1) - D(P\|P_\alpha), \tag{46}$$

where, thanks to the assumption that $D(P\,\|\,P_\alpha) < \infty$, we have invoked Corollary A1 in the Appendix twice with $(P, Q, R) \leftarrow (P, P_\alpha, P_0)$ and $(P, Q, R) \leftarrow (P, P_\alpha, P_1)$, respectively;

(2) $\max\{D(P\,\|\,P_0), D(P\,\|\,P_1)\} < \infty$: The proof is identical since we are entitled to invoke Corollary A1 to show (45) (resp., (46)) because $D(P\,\|\,P_0) < \infty$ (resp., $D(P\,\|\,P_1) < \infty$).

(3) $D(P\,\|\,P_\alpha) = \infty$ and $\max\{D(P\,\|\,P_0), D(P\,\|\,P_1)\} = \infty$: both sides of (43) are equal to ∞.

Finally, to show that (44) follows from (43), simply recall from (31) that $D_\alpha(P_1\,\|P_0) < \infty$. □

15. Relative entropy and Rényi divergence are related by the following fundamental variational representation.

Theorem 3. *Fix $\alpha \in (0,1)$ and $(P_1, P_0) \in \mathcal{P}_\mathcal{A}^2$. Then, the Rényi divergence between P_1 and P_0 satisfies*

$$(1-\alpha)\, D_\alpha(P_1\|P_0) = \min_P \{\alpha\, D(P\|P_1) + (1-\alpha)\, D(P\|P_0)\}, \tag{47}$$

where the minimum is over $\mathcal{P}_\mathcal{A}$. If $P_0 \not\perp P_1$, then the right side of (47) is attained by the tilted measure P_α, and the minimization can be restricted to the subset of probability measures which are dominated by both P_1 and P_0.

Proof. If $P_0 \perp P_1$, then both sides of (47) are $+\infty$ since there is no probability measure that is dominated by both P_0 and P_1. If $P_0 \not\perp P_1$, then minimizing both sides of (43) with respect to P yields (47) and the fact that the tilted probability measure attains the minimum therein. □

The variational representation in (47) was observed in [39] in the finite-alphabet case, and, contemporaneously, in full generality in [50]. Unlike Theorem 3, both of those references also deal with $\alpha > 1$. The function $d(\alpha) = (1-\alpha)\, D_\alpha(P_1\|P_0)$, with $d(1) = \lim_{\alpha\uparrow 1} d(\alpha)$, is concave in α because the right side of (47) is a minimum of affine functions of α.

16. Given random transformations $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$, $Q_{Y|X}\colon \mathcal{A} \to \mathcal{B}$, and a probability measure $P_X \in \mathcal{P}_\mathcal{A}$ on the input space, the conditional relative entropy is

$$D(P_{Y|X}\,\|\,Q_{Y|X}\,|\,P_X) = D(P_{Y|X}P_X\,\|\,Q_{Y|X}P_X) \tag{48}$$

$$= \mathbb{E}\Big[D\Big(P_{Y|X}(\cdot|X)\,\|\,Q_{Y|X}(\cdot|X)\Big)\Big], \quad X \sim P_X. \tag{49}$$

Analogously, the conditional Rényi divergence is defined as

$$D_\alpha(P_{Y|X}\,\|\,Q_{Y|X}\,|\,P_X) = D_\alpha(P_{Y|X}P_X\,\|\,Q_{Y|X}P_X). \tag{50}$$

A word of caution: the notation in (50) conforms to that in [38,45] but it is not universally adopted, e.g., [43] uses the left side of (50) to denote the Rényi generalization of the right side of (49). We can express the conditional Rényi divergence as

$$\begin{aligned} &D_\alpha(P_{Y|X} \,\|\, Q_{Y|X}|P_X) \\ &\quad = \frac{1}{\alpha-1}\log \mathbb{E}\Big[\exp\Big((\alpha-1)D_\alpha\Big(P_{Y|X}(\cdot|X) \,\|\, Q_{Y|X}(\cdot|X)\Big)\Big)\Big], \quad X \sim P_X, \qquad (51)\\ &\quad = \frac{1}{\alpha-1}\log \mathbb{E}\left[\left(\frac{\mathrm{d}P_{Y|X}}{\mathrm{d}Q_{Y|X}}(Y|X)\right)^{\alpha-1}\right], \quad (X,Y) \sim P_X P_{Y|X}, \qquad (52)\end{aligned}$$

where (52) holds if $P_X P_{Y|X} \ll P_X Q_{Y|X}$. Jensen's inequality applied to (51) results in

$$D_\alpha(P_{Y|X} \,\|\, Q_{Y|X}|P_X) \leqslant \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X) \,\|\, Q_{Y|X}(\cdot|X))\Big], \quad \alpha \in (0,1); \qquad (53)$$

$$D_\alpha(P_{Y|X} \,\|\, Q_{Y|X}|P_X) \geqslant \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X) \,\|\, Q_{Y|X}(\cdot|X))\Big], \quad \alpha > 1. \qquad (54)$$

Nevertheless, an immediate and crucial observation we can draw from (51) is that the unconstrained maximizations of the sides of (53) and of (54) over P_X do coincide: for all $\alpha > 0$,

$$\begin{aligned} \sup_X D_\alpha(P_{Y|X} \,\|\, Q_{Y|X}|P_X) &= \sup_X \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X) \,\|\, Q_{Y|X}(\cdot|X))\Big] \qquad (55)\\ &= \sup_{a \in \mathcal{A}} D_\alpha(P_{Y|X=a} \,\|\, Q_{Y|X=a}). \qquad (56)\end{aligned}$$

17. Conditional Rényi divergence satisfies the following additive decomposition, originally pointed out, without proof, by Sibson [31] in the setting of finite $\mathcal{A}$.

Theorem 4. *Given $P_X \in \mathcal{P}_\mathcal{A}$, $Q_Y \in \mathcal{P}_\mathcal{B}$, $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$, and $\alpha \in (0,1) \cup (1,\infty)$, we have*

$$D_\alpha(P_{Y|X} \,\|\, Q_Y|P_X) = D_\alpha(P_{Y|X} \,\|\, P_{Y[\alpha]}|P_X) + D_\alpha(P_{Y[\alpha]} \| Q_Y). \qquad (57)$$

Furthermore, with κ_α as in (14),

$$D_\alpha\Big(P_{Y|X} \,\|\, P_{Y[\alpha]}|P_X\Big) = \frac{\kappa_\alpha}{\alpha-1}. \qquad (58)$$

Proof. Select an arbitrary probability measure $R_Y \in \mathcal{P}_\mathcal{B}$ that dominates both Q_Y and P_Y, and, therefore, $P_{Y[\alpha]}$ too. Letting $(X,Z) \sim P_X \times R_Y$, we have

$$\begin{aligned} D_\alpha(P_{Y|X} \,\|\, Q_Y|P_X) &= \frac{1}{\alpha-1}\log \mathbb{E}\left[\left(\frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X \times R_Y}(X,Z)\right)^\alpha \left(\frac{\mathrm{d}Q_Y}{\mathrm{d}R_Y}(Z)\right)^{1-\alpha}\right] \qquad (59)\\ &= \frac{1}{\alpha-1}\log \mathbb{E}\left[\mathbb{E}\big[\exp\big(\alpha\, \imath_{X;Y}(X;Z)\big)|Z\big]\left(\frac{\mathrm{d}P_Y}{\mathrm{d}R_Y}(Z)\right)^\alpha \left(\frac{\mathrm{d}Q_Y}{\mathrm{d}R_Y}(Z)\right)^{1-\alpha}\right] \qquad (60)\\ &= \frac{\kappa_\alpha}{\alpha-1} + \frac{1}{\alpha-1}\log \mathbb{E}\left[\left(\frac{\mathrm{d}P_{Y[\alpha]}}{\mathrm{d}P_Y}(Z)\right)^\alpha \left(\frac{\mathrm{d}P_Y}{\mathrm{d}R_Y}(Z)\right)^\alpha \left(\frac{\mathrm{d}Q_Y}{\mathrm{d}R_Y}(Z)\right)^{1-\alpha}\right] \qquad (61)\\ &= \frac{\kappa_\alpha}{\alpha-1} + \frac{1}{\alpha-1}\log \mathbb{E}\left[\left(\frac{\mathrm{d}P_{Y[\alpha]}}{\mathrm{d}R_Y}(Z)\right)^\alpha \left(\frac{\mathrm{d}Q_Y}{\mathrm{d}R_Y}(Z)\right)^{1-\alpha}\right] \qquad (62)\\ &= \frac{\kappa_\alpha}{\alpha-1} + D_\alpha(P_{Y[\alpha]} \| Q_Y), \qquad (63)\end{aligned}$$

where (61) follows from (13), and (62) follows from the chain rule of Radon-Nikodym derivatives applied to $P_{Y[\alpha]} \ll P_Y \ll R_Y$. Then, (58) follows by specializing $Q_Y = P_{Y[\alpha]}$, and the proof of (57) is complete, upon plugging (58) into the right side of (63). □

A proof of (57) in the discrete case can be found in Appendix A of [37].

18. For all $\alpha > 0$, given two inputs $(P_X, Q_X) \in \mathcal{P}_{\mathcal{A}}^2$ and one random transformation $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$, Rényi divergence (and, in particular, relative entropy) satisfies the data processing inequality,

$$D_\alpha(P_X \,\|\, Q_X) \geqslant D_\alpha(P_Y \,\|\, Q_Y), \tag{64}$$

where $P_X \to P_{Y|X} \to P_Y$, and $Q_X \to P_{Y|X} \to Q_Y$. The data processing inequality for Rényi divergence was observed by Csiszár [52] in the more general context of f-divergences. More recently it was stated in [39,50]. Furthermore, given one input $P_X \in \mathcal{P}_{\mathcal{A}}$ and two transformations $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$ and $Q_{Y|X}\colon \mathcal{A} \to \mathcal{B}$, conditioning cannot decrease Rényi divergence,

$$D_\alpha(P_{Y|X} \,\|\, Q_{Y|X}|P_X) \geqslant D_\alpha(P_Y \,\|\, Q_Y). \tag{65}$$

Since $D_\alpha(P_{Y|X} \,\|\, Q_{Y|X}|P_X) = D_\alpha(P_X P_{Y|X} \,\|\, P_X Q_{Y|X})$, (65) follows by applying (64) to a deterministic transformation which takes an input pair and outputs the second component. Inequalities (53) and (65) imply the convexity of $D_\alpha(P\|Q)$ in (P, Q) for $\alpha \in (0, 1]$.

4. Dependence Measures

In this paper we are interested in three information measures that quantify the dependence between random variables X and Y, such that $P_X \to P_{Y|X} \to P_Y$, namely, mutual information, and two of its generalizations, α- mutual information and Augustin–Csiszár mutual information of order α.

19. The mutual information is

$$\begin{aligned} I(X;Y) = I(P_X, P_{Y|X}) &= D(P_{Y|X} \,\|\, P_Y \,|\, P_X) && (66)\\ &= \min_{Q_Y} D(P_{Y|X} \,\|\, Q_Y \,|\, P_X) && (67)\\ &= \min_{Q_Y} D(P_{XY} \,\|\, P_X \times Q_Y). && (68) \end{aligned}$$

20. Given $\alpha \in (0, 1) \cup (1, \infty)$, the α-mutual information is defined as (see [30–32,40,42,45])

$$\begin{aligned} I_\alpha(X;Y) &= I_\alpha(P_X, P_{Y|X}) && (69)\\ &= \min_{Q_Y} D_\alpha(P_{Y|X} \,\|\, Q_Y \,|\, P_X) && (70)\\ &= \min_{Q_Y} D_\alpha(P_{XY} \,\|\, P_X \times Q_Y) && (71)\\ &= D_\alpha\Big(P_{Y|X} \,\|\, P_{Y[\alpha]} \,|\, P_X\Big) && (72)\\ &= \frac{1}{\alpha - 1} \log \mathbb{E}\Big[\exp\Big((\alpha - 1) D_\alpha\Big(P_{Y|X}(\cdot|X) \,\|\, P_{Y[\alpha]}\Big)\Big)\Big], \quad X \sim P_X && (73)\\ &= D_\alpha\Big(P_{Y|X} \,\|\, P_Y|P_X\Big) - D_\alpha\Big(P_{Y[\alpha]} \,\|\, P_Y\Big) && (74)\\ &= \frac{\kappa_\alpha}{\alpha - 1} && (75)\\ &= \frac{\alpha}{\alpha - 1} \log \mathbb{E}[\mathbb{E}^{\frac{1}{\alpha}}[\exp(\alpha\, \imath_{X;Y}(X;\bar{Y})) \,|\, \bar{Y}]], \quad (X, \bar{Y}) \sim P_X \times P_Y, && (76) \end{aligned}$$

where (72) and (74) follow from (57); (73) is a special case of (51); (75) follows from Theorem 4; and, (76) is (14). In view of (67) and (69), we let $I_1(X;Y) = I(X;Y)$. The

notation we use for α-mutual information conforms to that used in [40,42,45,53]. Other notations include K_α in [32,38,39] and I_α^g in [43]. $I_0(X;Y)$ and $I_\infty(X;Y)$ are defined by taking the corresponding limits.

21. Theorem 4 and (72) result in the additive decomposition

$$I_\alpha(X;Y) = D_\alpha(P_{Y|X} \,\|\, Q_Y|P_X) - D_\alpha(P_{Y[\alpha]} \,\|\, Q_Y), \tag{77}$$

for any Q_Y with $D_\alpha(P_{Y[\alpha]} \,\|\, Q_Y) < \infty$, thereby generalizing the well-known decomposition for mutual information,

$$I(X;Y) = D(P_{Y|X} \,\|\, Q_Y|P_X) - D(P_Y \,\|\, Q_Y), \tag{78}$$

which, in contrast to (77), is a simple consequence of the chain rule whenever the dependence between X and Y is regular, and of Lemma A1 in general.

22. **Example 6.** *Additive independent Gaussian noise. If $Y = X + N$, where $X \sim \mathcal{N}(0, \sigma_X^2)$ independent of $N \sim \mathcal{N}(0, \sigma_N^2)$, then, for $\alpha > 0$,*

$$Y[\alpha] \;\sim \mathcal{N}\Big(0, \alpha\,\sigma_X^2 + \sigma_N^2\Big), \tag{79}$$

$$I_\alpha(X; X+N) = I_\alpha(X+N; X) = \frac{1}{2}\log\left(1 + \alpha\,\frac{\sigma_X^2}{\sigma_N^2}\right). \tag{80}$$

23. If $\alpha \in (0,1)$, (47) and (69) result in

$$\begin{aligned}&(1-\alpha) I_\alpha(P_X, P_{Y|X})\\ &\quad = \min_{Q_X Q_{Y|X}} \Big\{ D(Q_X \,\|\, P_X) + \alpha\, D(Q_{Y|X} \,\|\, P_{Y|X} \,|\, Q_X) + (1-\alpha)\, I(Q_X, Q_{Y|X}) \Big\}. \end{aligned} \tag{81}$$

For $\alpha > 1$ a proof of (81) is given in [39] for finite alphabets.

24. Unlike $I(P_X, P_{Y|X})$, we can express $I_\alpha(P_X, P_{Y|X})$ directly in terms of its arguments without involving the corresponding output distribution or the α-response to P_X. This is most evident in the case of discrete alphabets, in which (76) becomes

$$I_\alpha(X;Y) = \frac{\alpha}{\alpha-1} \log \sum_{y\in\mathcal{B}} \left(\sum_{x\in\mathcal{A}} P_X(x) P_{Y|X=x}^\alpha(y) \right)^{\frac{1}{\alpha}}, \tag{82}$$

$$I_0(X;Y) = -\log \max_{y\in\mathcal{B}} \sum_{x\in\mathcal{A}} P_X(x) 1\{P_{Y|X}(y|x) > 0\}, \tag{83}$$

$$I_\infty(X;Y) = \log \left(\sum_{b\in\mathcal{Y}} \sup_{a:\, P_X(a)>0} P_{Y|X}(b|a) \right). \tag{84}$$

For example, if X is discrete and $H_\alpha(X)$ denotes the Rényi entropy of order α, then for all $\alpha > 0$,

$$H_\alpha(X) = I_{\frac{1}{\alpha}}(X;X). \tag{85}$$

If X and Y are equiprobable with $\mathbb{P}[X \neq Y] = \delta$, then, in bits, $I_\alpha(X;Y) = 1 - h_\alpha(\delta)$, where $h_\alpha(\delta)$ denotes the binary Rényi entropy.

25. In the main region of interest, namely, $\alpha \in (0,1)$, frequently we use a different parametrization in terms of $\rho > 0$, with $\alpha = \frac{1}{1+\rho}$.

Theorem 5. *For any $\rho > 0$, we have the upper bound*

$$\rho\, I_{\frac{1}{1+\rho}}(X;Y) \leqslant \min_{Q_{Y|X}\colon \mathcal{A}\to\mathcal{B}} \Big\{ D(Q_{Y|X}\|P_{Y|X}\,|\,P_X) + \rho\, I(P_X, Q_{Y|X}) \Big\}. \tag{86}$$

Proof. Fix $Q_{Y|X}\colon \mathcal{A} \to \mathcal{B}$, and let $P_X \to Q_{Y|X} \to Q_Y$. Then,

$$\begin{aligned} I_{\frac{1}{1+\rho}}(X;Y) &\leqslant D_{\frac{1}{1+\rho}}(P_{XY}\|P_X \times Q_Y) && (87)\\ &= \frac{1+\rho}{\rho} \min_{R_{XY}} \left\{ \frac{1}{1+\rho} D(R_{XY}\|P_{XY}) + \frac{\rho}{1+\rho} D(R_{XY}\|P_X \times Q_Y) \right\} && (88)\\ &\leqslant \frac{1}{\rho}\, D(Q_{Y|X}P_X\|P_{XY}) + D(Q_{Y|X}P_X\|P_X \times Q_Y) && (89)\\ &= \frac{1}{\rho}\, D(Q_{Y|X}\|P_{Y|X}|P_X) + I(P_X, Q_{Y|X}), && (90) \end{aligned}$$

where (87), (88) and (90) follow from (69), (47) and (66) respectively. □

Just like (53), we will show in Section 7 that (86) becomes an equality upon the unconstrained maximization of both sides.

26. Before introducing the last dependence measure in this section, recall from Definition 7 and (58) that $P_{Y[\alpha]} \ll P_Y$, the α-response (of $P_{Y|X}$) to P_X defined by

$$\imath_{Y[\alpha]\|Y}(y) = \frac{1}{\alpha} \log \mathbb{E}[\exp\Big(\alpha\, \imath_{X;Y}(X;y) + (1-\alpha) D_\alpha\Big(P_{Y|X} \,\|\, P_{Y[\alpha]}|P_X\Big)\Big)], \tag{91}$$

attains $\min_{Q_Y} D_\alpha(P_{Y|X}\|Q_Y|P_X)$, where the expectation is with respect to $X \sim P_X$. We proceed to define $P_{Y\langle\alpha\rangle} \ll P_Y$, the $\langle\alpha\rangle$-response (of $P_{Y|X}$) to P_X by means of

$$\imath_{Y\langle\alpha\rangle\|Y}(y) = \frac{1}{\alpha} \log \mathbb{E}\Big[\exp(\alpha\, \imath_{X;Y}(X;y) + (1-\alpha) D_\alpha\Big(P_{Y|X}(\cdot|X) \,\|\, P_{Y\langle\alpha\rangle}\Big)\Big], \tag{92}$$

with $X \sim P_X$. Note that $P_{Y\langle 1\rangle} = P_{Y[1]} = P_Y$.

27. In the case of discrete alphabets, (92) becomes the implicit equation

$$P^\alpha_{Y\langle\alpha\rangle}(y) = \sum_{a\in\mathcal{A}} P_X(a) \frac{P^\alpha_{Y|X}(y|a)}{\sum_{b\in\mathcal{B}} P^\alpha_{Y|X}(b|a)\, P^{1-\alpha}_{Y\langle\alpha\rangle}(b)}, \quad y \in \mathcal{B}, \tag{93}$$

which coincides with (9.24) in Fano's 1961 textbook [7], with $s \leftarrow 1-\alpha$, and is also given by Haroutunian in (19) of [22]. For example, if $\mathcal{A} = \mathcal{B}$ is discrete and $Y = X$, then $P_{Y\langle\alpha\rangle} = P_X$, while $P^\alpha_{Y[\alpha]}(y) = c\, P_X(y)$, $y \in \mathcal{A}$.

28. The $\langle\alpha\rangle$-response satisfies the following identity, which can be regarded as the counterpart of (57) satisfied by the α-response.

Theorem 6. *Fix $P_X \in \mathcal{P}_\mathcal{A}$, $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$ and $Q_Y \in \mathcal{P}_\mathcal{B}$. Then,*

$$\begin{aligned} &D_\alpha(P_{Y\langle\alpha\rangle}\,\|\,Q_Y)\\ &\quad= \frac{1}{\alpha-1} \log \mathbb{E}\Big[\exp\Big((1-\alpha)\Big(D_\alpha(P_{Y|X}(\cdot|X)\|P_{Y\langle\alpha\rangle}) - D_\alpha(P_{Y|X}(\cdot|X)\|Q_Y)\Big)\Big)\Big]. \end{aligned} \tag{94}$$

Proof. For brevity we assume $Q_Y \ll P_Y$. Otherwise, the proof is similar adopting a reference measure that dominates both Q_Y and P_Y. The definition of unconditional

Rényi divergence in Item 11 implies that we can write $(\alpha-1)$ times the exponential of the left side of (94) as

$$\exp\Big((\alpha-1)D_\alpha(P_{Y\langle\alpha\rangle}\|Q_Y)\Big) = \mathbb{E}\left[\left(\frac{\mathrm{d}P_{Y\langle\alpha\rangle}}{\mathrm{d}P_Y}(Y)\right)^{\alpha}\left(\frac{\mathrm{d}Q_Y}{\mathrm{d}P_Y}(Y)\right)^{1-\alpha}\right] \tag{95}$$

$$= \mathbb{E}\left[\exp\Big(\alpha\, \imath_{X;Y}(X;Y) + (1-\alpha)D_\alpha\Big(P_{Y|X}(\cdot|X)\,\|\,P_{Y\langle\alpha\rangle}\Big)\Big)\left(\frac{\mathrm{d}Q_Y}{\mathrm{d}P_Y}(Y)\right)^{1-\alpha}\right] \tag{96}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\exp\Big(\alpha\, \imath_{X;Y}(X;Y) + (1-\alpha)\Big(\imath_{Q_Y\|P_Y}(Y) + D_\alpha\Big(P_{Y|X}(\cdot|X)\,\|\,P_{Y\langle\alpha\rangle}\Big)\Big)\Big)\Big|X\right]\right]$$

$$= \mathbb{E}\Big[\exp\Big((1-\alpha)\Big(D_\alpha\Big(P_{Y|X}(\cdot|X)\,\|\,P_{Y\langle\alpha\rangle}\Big) - D_\alpha\Big(P_{Y|X}(\cdot|X)\,\|\,Q_Y\Big)\Big)\Big)\Big], \tag{97}$$

where $(X,Y)\sim P_X\times P_Y$, (96) follows from (92), and (97) follows from the definition of unconditional Rényi divergence in (27). □

Theorem 7. *If $\alpha\in(0,1]$, then*

$$D_\alpha(P_{Y\langle\alpha\rangle}\,\|\,Q_Y) \leqslant \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X)\,\|\,Q_Y)\Big] - \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X)\,\|\,P_{Y\langle\alpha\rangle})\Big] \tag{98}$$

$$\leqslant D(P_{Y\langle\alpha\rangle}\,\|\,Q_Y). \tag{99}$$

If $\alpha\geqslant 1$, inequalities (98) and (99) are reversed.

Proof. Assume $\alpha\in(0,1]$. Jensen's inequality applied to the right side of (94) results in (98). To show (99), again we assume for brevity $Q_Y\ll P_Y$, and define the positive functions $V\colon\mathcal{A}\times\mathcal{B}\to(0,\infty)$ and $W\colon\mathcal{A}\times\mathcal{B}\to(0,\infty)$,

$$V(x,y) = \exp\Big(\alpha\imath_{X;Y}(x;y) + (1-\alpha)\imath_{Y\langle\alpha\rangle\|Y}(y)\Big), \tag{100}$$

$$W(x,y) = \exp\Big(\alpha\imath_{X;Y}(x;y) + (1-\alpha)\imath_{Q_Y\|P_Y}(y)\Big). \tag{101}$$

Note that, with $(X,Y)\sim P_X\times P_Y$, and $(x,y)\in\mathcal{A}\times\mathcal{B}$,

$$\mathbb{E}[V(x,Y)] = \exp\Big((\alpha-1)D_\alpha(P_{Y|X=x}\|P_{Y\langle\alpha\rangle})\Big), \tag{102}$$

$$\mathbb{E}[W(x,Y)] = \exp\Big((\alpha-1)D_\alpha(P_{Y|X=x}\|Q_Y)\Big), \tag{103}$$

$$\mathbb{E}\left[\frac{V(X,y)}{\mathbb{E}[V(X,Y)|X]}\right] = \exp\Big((1-\alpha)\imath_{Y\langle\alpha\rangle\|Y}(y)\Big)\cdot$$

$$\cdot\,\mathbb{E}\Big[\exp\Big(\alpha\,\imath_{X;Y}(X;y) + (1-\alpha)D_\alpha(P_{Y|X}(\cdot|X)\|P_{Y\langle\alpha\rangle})\Big)\Big] \tag{104}$$

$$= \frac{\mathrm{d}P_{Y\langle\alpha\rangle}}{\mathrm{d}P_Y}(y). \tag{105}$$

where (104) uses (100) and (102) and (105) follows from (92). Then,

$$D_\alpha(P_{Y|X=x}\,\|\,Q_Y) - D_\alpha(P_{Y|X=x}\,\|\,P_{Y\langle\alpha\rangle})$$

$$= \frac{1}{1-\alpha}\log\frac{\mathbb{E}[V(x,Y)]}{\mathbb{E}[W(x,Y)]} \tag{106}$$

$$\leqslant \frac{1}{1-\alpha}\mathbb{E}\left[\frac{V(x,Y)}{\mathbb{E}[V(x,Y)]}\log\frac{V(x,Y)}{W(x,Y)}\right] \tag{107}$$

$$= \mathbb{E}\left[\frac{V(x,Y)}{\mathbb{E}[V(x,Y)]}\Big(\imath_{Y\langle\alpha\rangle\|Y}(Y) - \imath_{Q_Y\|P_Y}(Y)\Big)\right], \tag{108}$$

where the expectations are with respect to $Y\sim P_Y$, and

- (107) follows from the log-sum inequality for integrable non-negative random variables,

$$\mathbb{E}[V]\log\frac{\mathbb{E}[V]}{\mathbb{E}[W]} \leqslant \mathbb{E}\left[V\log\frac{V}{W}\right]; \tag{109}$$

- (108) $\Leftarrow$ (100) and (101).

Taking expectation with respect to $X \sim P_X$ of (106)–(108) yields (99) because of Lemma A1 and (105). If $\alpha \geqslant 1$, then Jensen's inequality applied to the right side of (94) results in (98) but with the opposite inequality. Moreover, (107) is reversed and the remainder of the proof holds verbatim. □

In the case of finite input-alphabets, a different proof of (99) is given in Appendix B of [54].

29. Introduced in the unpublished dissertation [36] and rescued from oblivion in [32], the Augustin–Csiszár mutual information of order α is defined for $\alpha > 0$ as

$$I_\alpha^{\mathsf{c}}(X;Y) = I_\alpha^{\mathsf{c}}(P_X, P_{Y|X}) = \min_{Q_Y} \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X)\,\|\,Q_Y)\Big] \tag{110}$$

$$= \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X)\,\|\,P_{Y\langle\alpha\rangle})\Big], \tag{111}$$

where (111) follows from (98) if $\alpha \in (0,1]$, and from the reverse of (99) if $\alpha \geqslant 1$. We conform to the notation in [40], where I_α^{a} was used to denote the difference between entropy and Arimoto-Rényi conditional entropy. In [32,39,43] the Augustin–Csiszár mutual information of order α is denoted by I_α. In Augustin's original notation [36], $I^\rho(P_X)$ means $I^{\mathsf{c}}_{1-\rho}(P_X, P_{Y|X})$, $\rho \in (0,1)$. Independently of [36], Poltyrev [35] introduced a functional (expressed as a maximization over a reverse random transformation) which turns out to be $\rho I^{\mathsf{c}}_{\frac{1}{1+\rho}}(X;Y)$ and which he denoted by $E_0(\rho, P_X)$, although in Gallager's notation that corresponds to $\rho I_{\frac{1}{1+\rho}}(X;Y)$, as we will see in (233). $I_0^{\mathsf{c}}(X;Y)$ and $I_\infty^{\mathsf{c}}(X;Y)$ are defined by taking the corresponding limits.

30. In the discrete case, (110) boils down to

$$I_\alpha^{\mathsf{c}}(X;Y) = \min_{Q_Y} \frac{1}{\alpha-1}\sum_{x\in\mathcal{A}} P_X(x)\log\sum_{y\in\mathcal{B}} P_{Y|X}^\alpha(y|x)\,Q_Y^{1-\alpha}(y), \tag{112}$$

which can be juxtaposed with the much easier expression in (82) for $I_\alpha(X;Y)$ involving no further optimization. Minimizing the Lagrangian, we can verify that the minimizer in (112) satisfies (93). With $(X,\bar{Y}) \sim P_X \times Q_Y$, we have

$$I_0^{\mathsf{c}}(X;Y) = \min_{Q_Y} \mathbb{E}\left[\log\frac{1}{\mathbb{P}[P_{Y|X}(\bar{Y}|X) > 0 \mid X]}\right], \tag{113}$$

$$I_\infty^{\mathsf{c}}(X;Y) = \min_{Q_Y} \mathbb{E}\left[\log\left\|\frac{P_{Y|X}(\bar{Y}|X)}{Q_Y(\bar{Y})}\right\|_\infty\right], \tag{114}$$

where the expectations are with respect to X.

31. The respective minimizers of (72) and (110), namely, the α-response and the $\langle\alpha\rangle$-response, are quite different. Most notably, in contrast to Item 7, an explicit expression for $P_{Y\langle\alpha\rangle}$ is unknown. Instead of defining $P_{Y\langle\alpha\rangle}$ through (92), [36] defines it, equivalently, as the fixed point of the operator (dubbed the Augustin operator in [43]) which maps the set of probability measures on the output space to itself,

$$\frac{\mathrm{d}\mathbb{T}_\alpha(Q)}{\mathrm{d}Q}(y) = \mathbb{E}\left[\left(\frac{\mathrm{d}P_{Y|X}}{\mathrm{d}Q}(y|X)\right)^\alpha \exp\Big((1-\alpha)D_\alpha(P_{Y|X}(\cdot|X)\|Q)\Big)\right], \tag{115}$$

where $X \sim P_X$. Although we do not rely on them, Lemma 34.2 of ($\alpha \in (0,1)$) and Lemma 13 of [43] ($\alpha > 1$) claim that the minimizer in (110), referred to in [43] as the Augustin mean of order α, is unique and is a fixed point of the operator $\mathbb{T}_\alpha$ regardless of P_X. Moreover, Lemma 13(c) of [43] establishes that for $\alpha \in (0,1)$ and finite input alphabets, repeated iterations of the operator $\mathbb{T}_\alpha$ with initial argument $P_{Y[\alpha]}$ converge to $P_{Y\langle\alpha\rangle}$.

32. It is interesting to contrast the next example with the formulas in Examples 2 and 6.

Example 7. *Additive independent Gaussian noise. If $Y = X + N$, where $X \sim \mathcal{N}(0, \sigma_X^2)$ independent of $N \sim \mathcal{N}(0, \sigma_N^2)$, then*

$$Y\langle\alpha\rangle \sim \mathcal{N}\left(0, \frac{\sigma_N^2}{2}\left(2 - \frac{1}{\alpha} + \Delta + \mathsf{snr}\right)\right), \tag{116}$$

$$\mathsf{snr} = \frac{\sigma_X^2}{\sigma_N^2}, \tag{117}$$

$$\Delta = \sqrt{4\,\mathsf{snr} + \left(\frac{1}{\alpha} - \mathsf{snr}\right)^2}. \tag{118}$$

This result can be obtained by postulating a zero-mean Gaussian distribution with variance v_α^2 as $P_{Y\langle\alpha\rangle}$ and verifying that (92) is indeed satisfied if v_α^2 is chosen as in (116). The first step is to invoke (32), which yields

$$D_\alpha\left(P_{Y|X=x} \,\|\, P_{Y\langle\alpha\rangle}\right) = \frac{\lambda_\alpha}{2} + \frac{\alpha\, x^2}{2\, s_\alpha^2}\log \mathrm{e}, \tag{119}$$

$$\lambda_\alpha = \log\frac{v_\alpha^2}{\sigma_N^2} + \frac{1}{\alpha - 1}\log\frac{v_\alpha^2}{s_\alpha^2}, \tag{120}$$

where we have denoted $s_\alpha^2 = \alpha\, v_\alpha^2 + (1-\alpha)\sigma_N^2$. Since $Y \sim \mathcal{N}(0, \sigma_X^2 + \sigma_N^2)$,

$$\imath_{X;Y}(x;y) = \frac{1}{2}\log\frac{\sigma_X^2 + \sigma_N^2}{\sigma_N^2} + \frac{1}{2}\left(\frac{y^2}{\sigma_X^2 + \sigma_N^2} - \frac{(y-x)^2}{\sigma_N^2}\right)\log \mathrm{e}, \tag{121}$$

$$\imath_{Y\langle\alpha\rangle\|Y}(y) = \frac{1}{2}\log\frac{\sigma_X^2 + \sigma_N^2}{v_\alpha^2} + \frac{1}{2}\left(\frac{y^2}{\sigma_X^2 + \sigma_N^2} - \frac{y^2}{v_\alpha^2}\right)\log \mathrm{e}. \tag{122}$$

Assembling (120) and (121), the right side of (92) becomes

$$\begin{aligned}
&\frac{1}{\alpha}\log\mathbb{E}\Big[\exp\big(\alpha\,\imath_{X;Y}(X;y) + (1-\alpha)D_\alpha\left(P_{Y|X}(\cdot|X) \,\|\, P_{Y\langle\alpha\rangle}\right)\big)\Big] \\
&= \frac{1}{2}\log\frac{\sigma_X^2 + \sigma_N^2}{\sigma_N^2} + \frac{1}{2}\frac{y^2\log \mathrm{e}}{\sigma_X^2 + \sigma_N^2} + \frac{1-\alpha}{2\alpha}\lambda_\alpha \\
&\quad + \frac{1}{\alpha}\log\mathbb{E}\left[\exp_{\mathrm{e}}\left(-\frac{\alpha(y-X)^2}{2\sigma_N^2} + \frac{\alpha(1-\alpha)X^2}{2s_\alpha^2}\right)\right] &(123)\\
&= \frac{1}{2}\log\frac{\sigma_X^2 + \sigma_N^2}{\sigma_N^2} + \frac{1-\alpha}{2\alpha}\lambda_\alpha + \frac{y^2\log \mathrm{e}}{2}\left(\frac{1}{\sigma_X^2 + \sigma_N^2} - \frac{s_\alpha^2 - \alpha(1-\alpha)\sigma_X^2}{\sigma_N^2 s_\alpha^2 + \alpha^2 v_\alpha^2\sigma_X^2}\right) \\
&\quad + \frac{1}{2\alpha}\log\frac{\sigma_N^2 s_\alpha^2}{\sigma_N^2 s_\alpha^2 + \alpha^2 v_\alpha^2\sigma_X^2} &(124)\\
&= \frac{1}{2}\log\frac{\sigma_X^2 + \sigma_N^2}{v_\alpha^2} + \frac{1}{2}\left(\frac{y^2}{\sigma_X^2 + \sigma_N^2} - \frac{y^2}{v_\alpha^2}\right)\log \mathrm{e}, &(125)
\end{aligned}$$

where (124) *follows by Gaussian integration, and the marvelous simplification in* (125) *is satisfied provided that we choose*

$$s_\alpha^2 = \frac{\alpha\,\sigma_X^2\,v_\alpha^2}{v_\alpha^2 - \sigma_N^2}. \tag{126}$$

Comparing (122) *and* (125)*, we see that* (92) *is indeed satisfied with* $Y\langle\alpha\rangle \sim \mathcal{N}(0, v_\alpha^2)$ *if* v_α^2 *satisfies the quadratic equation* (126)*, whose solution is in* (116)–(118)*. Invoking* (32) *and* (116)*, we obtain*

$$\begin{aligned} I_\alpha^{\mathsf{c}}(X; X+N) = {} & \frac{\alpha\,\mathsf{snr}}{1+\alpha\,\Delta+\alpha\,\mathsf{snr}}\log \mathrm{e} + \frac{1}{2}\log\left(1+\frac{1}{2}\left(\Delta+\mathsf{snr}-\frac{1}{\alpha}\right)\right) \\ & - \frac{1}{2(1-\alpha)}\log\frac{2-\frac{1}{\alpha}+\Delta+\mathsf{snr}}{1+\alpha\,\Delta+\alpha\,\mathsf{snr}}. \end{aligned} \tag{127}$$

Beyond its role in evaluating the Augustin–Csiszár mutual information for Gaussian inputs, the Gaussian distribution in (116) has found some utility in the analysis of finite blocklength fundamental limits for data transmission [55].

33. This item gives a variational representation for the Augustin–Csiszár mutual information in terms of mutual information and conditional relative entropy (i.e., non-Rényi information measures). As we will see in Section 10, this representation accounts for the role played by Augustin–Csiszár mutual information in expressing error exponent functions.

Theorem 8. *For* $\alpha \in (0,1)$*, the Augustin–Csiszár mutual information satisfies the variational representation in terms of conditional relative entropy and mutual information,*

$$(1-\alpha)\,I_\alpha^{\mathsf{c}}(P_X, P_{Y|X}) = \min_{Q_{Y|X}}\Big\{\alpha\,D(Q_{Y|X}\|P_{Y|X}|P_X) + (1-\alpha)\,I(P_X, Q_{Y|X})\Big\}, \tag{128}$$

where the minimum is over all the random transformations from the input to the output spaces.

Proof. Invoking (47) with $(P_1, P_0) \leftarrow (P_{Y|X=x}, Q_Y)$ we obtain

$$(1-\alpha)\,D_\alpha(P_{Y|X=x}\|Q_Y) = \min_{R_Y}\Big\{\alpha\,D(R_Y\|P_{Y|X=x}) + (1-\alpha)\,D(R_Y\|Q_Y)\Big\} \tag{129}$$

$$= \min_{R_{Y|X=x}}\Big\{\alpha\,D(R_{Y|X=x}\|P_{Y|X=x}) + (1-\alpha)\,D(R_{Y|X=x}\|Q_Y)\Big\}. \tag{130}$$

Averaging over $x \sim P_X$, followed by minimization with respect to Q_Y yields (128) upon recalling (67). □

In the finite-alphabet case with $\alpha \in (0,1)\cup(1,\infty)$, the representation in (128) is implicit in the appendix of [32], and stated explicitly in [39], where it is shown by means of a minimax theorem. This is one of the instances in which the proof of the result is considerably easier for $\alpha \in (0,1)$; we can take the following route to show (128) for $\alpha > 1$. Neglecting to emphasize its dependence on P_X, denote

$$f_\alpha(Q_Y, R_{Y|X}) = \frac{\alpha}{1-\alpha}\,D(R_{Y|X}\|P_{Y|X}|P_X) + D(R_{Y|X}\|Q_Y|P_X). \tag{131}$$

Invoking (47) we obtain

$$D_\alpha(P_{Y|X=x}\|Q_Y) = \max_{R_{Y|X=x}}\left\{\frac{\alpha}{1-\alpha}\,D(R_{Y|X=x}\|P_{Y|X=x}) + D(R_{Y|X=x}\|Q_Y)\right\}. \tag{132}$$

Averaging (132) with respect to P_X followed by minimization over Q_Y, results in

$$I_\alpha^{\mathsf{c}}(P_X, P_{Y|X}) = \min_{Q_Y} \max_{R_{Y|X}} f_\alpha(Q_Y, R_{Y|X}) \tag{133}$$

$$\geqslant \max_{R_{Y|X}} \min_{Q_Y} f_\alpha(Q_Y, R_{Y|X}) \tag{134}$$

$$= \max_{R_{Y|X}} \left\{ \frac{\alpha}{1-\alpha} D(R_{Y|X} \| P_{Y|X} | P_X) + I(P_X, R_{Y|X}) \right\}, \tag{135}$$

which shows $\geqslant$ in (128). If a minimax theorem can be invoked to show equality in (134), then (128) is established for $\alpha > 1$. For that purpose, for fixed $R_{Y|X}$, $f(\cdot, R_{Y|X})$ is convex and lower semicontinuous in Q_Y on the set where it is finite. Rewriting

$$f(Q_Y, R_{Y|X}) = \frac{1}{1-\alpha} D(R_{Y|X} \| P_{Y|X} | P_X) + D(R_{Y|X} \| Q_Y | P_X) - D(R_{Y|X} \| P_{Y|X} | P_X), \tag{136}$$

it can be seen that $f(Q_Y, \cdot)$ is upper semicontinuous and concave (if $\alpha > 1$). A different, and considerably more intricate route is taken in Lemma 13(d) of [43], which also gives (128) for $\alpha > 1$ assuming finite input alphabets.

34. Unlike mutual information, neither $I_\alpha(X;Y) = I_\alpha(Y;X)$ nor $I_\alpha^{\mathsf{c}}(X;Y) = I_\alpha^{\mathsf{c}}(Y;X)$ hold in general.

Example 8. *Erasure transformation. Let* $\mathcal{A} = \{0, 1\}, \mathcal{B} = \{0, 1, \mathsf{e}\}$,

$$P_{Y|X}(b|a) = \begin{cases} 1-\delta, & a = b; \\ \delta, & b = \mathsf{e}; \\ 0, & a \neq b \neq \mathsf{e}, \end{cases} \tag{137}$$

with $\delta \in (0,1)$, *and* $P_X(0) = \frac{1}{2}$. *Then, we obtain, for* $\alpha \in (0,1) \cup (1, \infty)$,

$$I_\alpha(X;Y) = I_\alpha^{\mathsf{c}}(X;Y) = \frac{\alpha}{\alpha - 1} \log\left(\delta + (1-\delta) 2^{\left(1 - \frac{1}{\alpha}\right)}\right), \tag{138}$$

$$I_\alpha(Y;X) = \frac{1}{\alpha - 1} \log\left(\delta + (1-\delta) 2^{\alpha - 1}\right), \tag{139}$$

$$I_\alpha^{\mathsf{c}}(Y;X) = I(X;Y) = 1 - \delta \text{ bits}. \tag{140}$$

35. It was shown in Theorem 5.2 of [38] that α-mutual information satisfies the data processing lemma, namely, if X and Z are conditionally independent given Y, then

$$I_\alpha(X;Z) \leqslant \min\{I_\alpha(X;Y), I_\alpha(Y;Z)\}, \tag{141}$$

$$I_\alpha(Z;X) \leqslant \min\{I_\alpha(Z;Y), I_\alpha(Y;X)\}. \tag{142}$$

As shown by Csiszár [32] using the data processing inequality for Rényi divergence, the data processing lemma also holds for I_α^{c}.

36. From (53), (54) and the monotonicity of $D_\alpha(P\|Q)$ in α, we obtain the ordering

$$I_\beta(X;Y) \leqslant I_\alpha(X;Y) \leqslant I_\alpha^{\mathsf{c}}(X;Y) \leqslant I_\nu^{\mathsf{c}}(X;Y) \leqslant I(X;Y), \quad 0 < \beta \leqslant \alpha \leqslant \nu < 1; \tag{143}$$

$$I(X;Y) \leqslant I_\nu^{\mathsf{c}}(X;Y) \leqslant I_\alpha^{\mathsf{c}}(X;Y) \leqslant I_\alpha(X;Y) \leqslant I_\beta(X;Y), \quad 1 < \nu \leqslant \alpha \leqslant \beta. \tag{144}$$

37. The convexity/concavity properties of the generalized mutual informations are summarized next.

Theorem 9.

(a) $\rho\, I_{\frac{1}{1+\rho}}(X;Y)$ *and* $\rho\, I^{\mathsf{c}}_{\frac{1}{1+\rho}}(X;Y)$ *are concave and monotonically non-decreasing in* $\rho \geqslant 0$.

(b) $I(\cdot, P_{Y|X})$ *and* $I^{\mathsf{c}}_{\alpha}(\cdot, P_{Y|X})$ *are concave functions. The same holds for* $I_\alpha(\cdot, P_{Y|X})$ *if* $\alpha > 1$.

(c) *If* $\alpha \in (0,1)$, *then* $I(P_X, \cdot)$, $I_\alpha(P_X, \cdot)$ *and* $I^{\mathsf{c}}_{\alpha}(P_X, \cdot)$ *are convex functions.*

Proof.

(a) According to (81) and (128), respectively, with $\alpha = \frac{1}{1+\rho} \in (0,1)$, $\rho\, I_{\frac{1}{1+\rho}}(X;Y)$ and $\rho\, I^{\mathsf{c}}_{\frac{1}{1+\rho}}(X;Y)$ are the infima of affine functions with nonnegative slopes.

(b) For mutual information the result goes back to [56] in the finite-alphabet case. In general, it holds since (67) is the infimum of linear functions of P_X. The same reasoning applies to Augustin–Csiszár mutual information in view of (110). For α-mutual information with $\alpha > 1$, notice from (51) that $D_\alpha(P_{Y|X} \,\|\, Q_Y|P_X)$ is concave in P_X if $\alpha > 1$. Therefore,

$$I_\alpha(\lambda\, P^1_X + (1-\lambda)\, P^0_X, P_{Y|X}) \tag{145}$$

$$= \inf_{Q_Y} D_\alpha(P_{Y|X} \,\|\, Q_Y|\, \lambda\, P^1_X + (1-\lambda)\, P^0_X) \tag{146}$$

$$\geqslant \inf_{Q_Y} \lambda\, D_\alpha(P_{Y|X} \,\|\, Q_Y|\, P^1_X) + (1-\lambda)\, D_\alpha(P_{Y|X} \,\|\, Q_Y|\, P^0_X) \tag{147}$$

$$\geqslant \lambda\, I_\alpha(P^1_X, P_{Y|X}) + (1-\lambda)\, I_\alpha(P^0_X, P_{Y|X}). \tag{148}$$

(c) The convexity of $I(P_X, \cdot)$ and $I_\alpha(P_X, \cdot)$ follow from the convexity of $D_\alpha(P\|Q)$ in (P,Q) for $\alpha \in (0,1]$ as we saw in Item 18. To show convexity of $I^{\mathsf{c}}_{\alpha}(P_X, \cdot)$ if $\alpha \in (0,1)$, we apply (169) in Item 45 with $P_{Y|X} = \lambda P^1_{Y|X} + (1-\lambda) P^0_{Y|X}$, and invoke the convexity of $I_\alpha(P_X, \cdot)$:

$$(1-\alpha)\, I^{\mathsf{c}}_{\alpha}(P_X, P_{Y|X})$$

$$= \max_{Q_X} \Big\{ (1-\alpha)\, I_\alpha(Q_X, \lambda P^1_{Y|X} + (1-\lambda) P^0_{Y|X}) - D(P_X \,\|\, Q_X) \Big\}, \tag{149}$$

$$\leqslant \max_{Q_X} \Big\{ \lambda \Big(1-\alpha)\, I_\alpha(Q_X, P^1_{Y|X}) - D(P_X \,\|\, Q_X) \Big)$$

$$+ (1-\lambda) \Big(1-\alpha)\, I_\alpha(Q_X, P^0_{Y|X}) - D(P_X \,\|\, Q_X) \Big) \Big\} \tag{150}$$

$$\leqslant (1-\alpha) \Big(\lambda I^{\mathsf{c}}_{\alpha}(P_X, P^1_{Y|X}) + (1-\lambda) I^{\mathsf{c}}_{\alpha}(P_X, P^0_{Y|X}) \Big). \tag{151}$$

□

Although not used in the sequel, we note, for completeness, that if $\alpha \in (0,1) \cup (1,\infty)$, [38] (see corrected version in [41]) shows that $\exp\Big(\Big(1 - \frac{1}{\alpha} \Big) I_\alpha(\cdot, P_{Y|X}) \Big) / (\alpha - 1)$ is concave.

5. Interplay between $I_\alpha(P_X, P_{Y|X})$ and $I^{\mathsf{c}}_{\alpha}(P_X, P_{Y|X})$

In this section we study the interplay between both notions of mutual informations of order α, and, in particular, various variational representations of these information measures.

38. For given $\alpha \in (0,1) \cup (1,\infty)$ and $P_{Y|X} \colon \mathcal{A} \to \mathcal{B}$, define $Q_{X[\alpha]} \ll\gg P_X$, the α-adjunct of P_X by

$$\imath_{Q_{X[\alpha]} \| P_X}(x) = (\alpha - 1)\, D_\alpha \Big(P_{Y|X=x} \| P_{Y[\alpha]} \Big) - \kappa_\alpha, \tag{152}$$

with κ_α the constant in (14) and $P_{Y[\alpha]}$, the α-response to P_X.

39. **Example 9.** *Let* $Y = X + N$ *with* $X \sim \mathcal{N}(0, \sigma_X^2)$ *independent of* $N \sim \mathcal{N}(0, \sigma_N^2)$, *and* $\mathsf{snr} = \frac{\sigma_X^2}{\sigma_N^2}$. *The* α*-adjunct of the input is*

$$Q_{X[\alpha]} = \mathcal{N}\left(0, \sigma_X^2 \frac{1 + \alpha^2\,\mathsf{snr}}{1 + \alpha\,\mathsf{snr}}\right). \tag{153}$$

40. **Theorem 10.** *The* $\langle\alpha\rangle$*-response to* $Q_{X[\alpha]}$ *is* $P_{Y[\alpha]}$, *the* α*-response to* P_X.

Proof. We just need to verify that (92) is satisfied if we substitute $Y\langle\alpha\rangle$ by $Y[\alpha]$, and instead of taking the expectation in the right side with respect to $X \sim P_X$ we take it with respect to $\widetilde{X} \sim Q_{X[\alpha]}$. Then,

$$\begin{aligned}
&\mathbb{E}\Big[\exp(\alpha\, \imath_{X;Y}(\widetilde{X}; y) + (1-\alpha) D_\alpha\Big(P_{Y|X}(\cdot|\widetilde{X}) \,\|\, P_{Y[\alpha]}\Big)\Big] \\
&= \mathbb{E}\Big[\exp\Big(\imath_{Q_{X[\alpha]}\|P_X}(X) + \alpha\, \imath_{X;Y}(X; y) + (1-\alpha) D_\alpha\Big(P_{Y|X}(\cdot|X) \,\|\, P_{Y[\alpha]}\Big)\Big)\Big] && (154)\\
&= \mathbb{E}\big[\exp(\alpha\, \imath_{X;Y}(X; y) - \kappa_\alpha)\big] && (155)\\
&= \exp\Big(\alpha\, \imath_{Y[\alpha]\|Y}(y)\Big), && (156)
\end{aligned}$$

where (154) is by change of measure, (155) follows by substitution of (152), and (156) is the same as (13). □

41. For given $\alpha \in (0,1) \cup (1,\infty)$ and $P_{Y|X} \colon \mathcal{A} \to \mathcal{B}$, we define $Q_{X\langle\alpha\rangle} \ll P_X$, the $\langle\alpha\rangle$-adjunct of an input probability measure P_X through

$$\imath_{Q_{X\langle\alpha\rangle}\|P_X}(x) = (1-\alpha)\, D_\alpha\Big(P_{Y|X=x} \,\|\, P_{Y\langle\alpha\rangle}\Big) + v_\alpha, \tag{157}$$

where $P_{Y\langle\alpha\rangle}$ is the $\langle\alpha\rangle$-response to P_X and v_α is a normalizing constant so that $Q_{X\langle\alpha\rangle}$ is a probability measure. According to (9), we must have

$$\mathbb{E}\Big[\exp\Big(\imath_{Q_{X\langle\alpha\rangle}\|P_X}(X)\Big)\Big] = 1, \quad X \sim P_X. \tag{158}$$

Hence,

$$v_\alpha = (\alpha - 1)\, D_\alpha\Big(P_{Y|X} \,\|\, P_{Y\langle\alpha\rangle} \,|\, Q_{X\langle\alpha\rangle}\Big). \tag{159}$$

42. With the aid of the expression in Example 7, we obtain

Example 10. *Let* $Y = X + N$ *with* $X \sim \mathcal{N}(0, \sigma_X^2)$ *independent of* $N \sim \mathcal{N}(0, \sigma_N^2)$, *and* $\mathsf{snr} = \frac{\sigma_X^2}{\sigma_N^2}$. *Then, the* $\langle\alpha\rangle$*-adjunct of the input is*

$$Q_{X\langle\alpha\rangle} = \mathcal{N}\left(0, \sigma_X^2\, \frac{1 + \alpha(\Delta + \mathsf{snr})}{1 + \alpha(\Delta - \mathsf{snr}) + 2\,\alpha^2\,\mathsf{snr}}\right), \tag{160}$$

which, in contrast to $Q_{X[\alpha]}$, *has larger variance than* σ_X^2 *if* $\alpha \in (0,1)$.

43. The following result is the dual of Theorem 10.

Theorem 11. *The* α*-response to* $Q_{X\langle\alpha\rangle}$ *is* $P_{Y\langle\alpha\rangle}$, *the* $\langle\alpha\rangle$*-response to* P_X. *Therefore,*

$$v_\alpha = (\alpha - 1)\, I_\alpha\Big(Q_{X\langle\alpha\rangle}, P_{Y|X}\Big). \tag{161}$$

Proof. The proof is similar to that of Theorem 10. We just need to verify that we obtain the right side of (92) if on the right side of (91) we substitute P_X by $Q_{X\langle\alpha\rangle}$ and $P_{Y[\alpha]}$ by $P_{Y\langle\alpha\rangle}$. Let $\bar{X} \sim Q_{X\langle\alpha\rangle}$. Then,

$$\begin{aligned}
&\frac{1}{\alpha}\log\mathbb{E}\Big[\exp\Big(\alpha\,\imath_{X;Y}(\bar{X};y)+(1-\alpha)D_\alpha\Big(P_{Y|X}\,\|\,P_{Y\langle\alpha\rangle}|Q_{X\langle\alpha\rangle}\Big)\Big)\Big]\\
&=\frac{1}{\alpha}\log\mathbb{E}\Big[\exp\Big(\imath_{Q_{X\langle\alpha\rangle}\|P_X}(X)+\alpha\,\imath_{X;Y}(X;y)-\upsilon_\alpha\Big)\Big] &(162)\\
&=\frac{1}{\alpha}\log\mathbb{E}\Big[\exp\Big(\alpha\,\imath_{X;Y}(X;y)+(1-\alpha)D_\alpha\Big(P_{Y|X}(\cdot|X)\,\|\,P_{Y\langle\alpha\rangle}\Big)\Big)\Big] &(163)\\
&=\imath_{Y\langle\alpha\rangle\|Y}(y), &(164)
\end{aligned}$$

where (162)–(164) follow by change of measure, (157), and (92), respectively. □

44. By recourse to a minimax theorem, the following representation is given for $\alpha \in (0,1)\cup(1,\infty)$ in the case of finite alphabets in [39], and dropping the restriction on the finiteness of the output space in [43]. As we show, a very simple and general proof is possible for $\alpha \in (0,1)$.

Theorem 12. *Fix $\alpha \in (0,1)$, $P_X \in \mathcal{P}_\mathcal{A}$ and $P_{Y|X}\colon \mathcal{A}\to\mathcal{B}$. Then,*

$$(1-\alpha)\,I_\alpha(X;Y)=\min_{Q_X}\Big\{(1-\alpha)\,I_\alpha^{\mathsf{c}}(Q_X,P_{Y|X})+D(Q_X\,\|\,P_X)\Big\}, \tag{165}$$

where the minimum is attained by $Q_{X[\alpha]}$, the α-adjunct of P_X defined in (152).

Proof. The variational representations in (81) and (128) result in (165). To show that the minimum is indeed attained by $Q_{X[\alpha]}$, recall from Theorem 10 that the $\langle\alpha\rangle$-response to $Q_{X[\alpha]}$ is $P_{Y[\alpha]}$. Therefore, evaluating the term in $\{\}$ in (165) for $Q_X \leftarrow Q_{X[\alpha]}$ yields, with $\widetilde{X}\sim Q_{X[\alpha]}$,

$$\begin{aligned}
&(1-\alpha)\,I_\alpha^{\mathsf{c}}(Q_{X[\alpha]},P_{Y|X})+D(Q_{X[\alpha]}\,\|\,P_X)\\
&\quad=(1-\alpha)\,\mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|\widetilde{X})\,\|\,P_{Y[\alpha]})\Big]+D(Q_{X[\alpha]}\,\|\,P_X) &(166)\\
&\quad=-\kappa_\alpha &(167)\\
&\quad=(1-\alpha)\,I_\alpha(X;Y), &(168)
\end{aligned}$$

where (167) follows from (152) and (168) results from (69)–(75). □

45. For finite-input alphabets, Lemma 18(b) of [43] (earlier Theorem 3.4 of [35] gave an equivalent variational characterization assuming, in addition, finite output alphabets) established the following dual to Theorem 12.

Theorem 13. *Fix $\alpha \in (0,1)$, $P_X \in \mathcal{P}_\mathcal{A}$ and $P_{Y|X}\colon \mathcal{A}\to\mathcal{B}$. Then,*

$$(1-\alpha)\,I_\alpha^{\mathsf{c}}(X;Y)=\max_{Q_X}\Big\{(1-\alpha)\,I_\alpha(Q_X,P_{Y|X})-D(P_X\,\|\,Q_X)\Big\}. \tag{169}$$

The maximum is attained by $Q_{X\langle\alpha\rangle}$, the $\langle\alpha\rangle$-adjunct of P_X defined by (157).

Proof. First observe that (165) implies that $\geqslant$ holds in (169). Second, the term in $\{\}$ on the right side of (169) evaluated at $Q_X \leftarrow Q_{X\langle\alpha\rangle}$ becomes

$$\begin{aligned}
&(1-\alpha)\,I_\alpha(Q_{X\langle\alpha\rangle},P_{Y|X})-D(P_X\,\|\,Q_{X\langle\alpha\rangle})\\
&\quad=(1-\alpha)\,I_\alpha(Q_{X\langle\alpha\rangle},P_{Y|X})+(1-\alpha)I_\alpha^{\mathsf{c}}(P_X,P_{Y|X})+\upsilon_\alpha &(170)\\
&\quad=(1-\alpha)I_\alpha^{\mathsf{c}}(P_X,P_{Y|X}), &(171)
\end{aligned}$$

where (170) follows by taking the expectation of minus (157) with respect to P_X. Therefore, $\leqslant$ also holds in (169) and the maximum is attained by $Q_{X\langle\alpha\rangle}$, as we wanted to show. □

Hinging on Theorem 8, Theorems 12 and 13 are given for $\alpha \in (0,1)$ which is the region of interest in the analysis of error exponents. Whenever, as in the finite-alphabet case, (128) holds for $\alpha > 1$, Theorems 12 and 13 also hold for $\alpha > 1$.
Notice that since the definition of $Q_{X\langle\alpha\rangle}$ involves $P_{Y\langle\alpha\rangle}$, the fact that it attains the maximum in (169) does not bring us any closer to finding $I_\alpha^{\mathsf{c}}(X;Y)$ for a specific input probability measure P_X. Fortunately, as we will see in Section 8, (169) proves to be the gateway to the maximization of $I_\alpha^{\mathsf{c}}(X;Y)$ in the presence of input-cost constraints.

46. Focusing on the main range of interest, $\alpha \in (0,1)$, we can express (169) as

$$I_\alpha^{\mathsf{c}}(P_X, P_{Y|X}) = \max_{Q_X} \left\{ I_\alpha(Q_X, P_{Y|X}) - \frac{1}{1-\alpha} D(P_X \,\|\, Q_X) \right\} \tag{172}$$

$$= \max_{\xi \geqslant 0} \left\{ \mathbb{I}(\xi) - \frac{\xi}{1-\alpha} \right\} \tag{173}$$

$$= \mathbb{I}(\xi_\alpha) - \frac{\xi_\alpha}{1-\alpha}, \tag{174}$$

where we have defined the function (dependent on α, P_X, and $P_{Y|X}$)

$$\mathbb{I}(\xi) = \max_{\substack{Q_X: \\ D(P_X\|Q_X) \leqslant \xi}} I_\alpha(Q_X, P_{Y|X}), \tag{175}$$

and ξ_α is the solution to

$$\dot{\mathbb{I}}(\xi_\alpha) = \frac{1}{1-\alpha}. \tag{176}$$

Recall that the maxima over the input distribution in (172) and (175) are attained by the $\langle\alpha\rangle$-adjunct $Q_{X\langle\alpha\rangle}$ defined in Item 41.

47. At this point it is convenient to summarize the notions of input and output probability measures that we have defined for a given α, random transformation $P_{Y|X}$, and input probability measure P_X:
    - P_Y: The familiar output probability measure $P_X \to P_{Y|X} \to P_Y$, defined in Item 5.
    - $P_{Y[\alpha]}$: The α-response to P_X, defined in Item 7. It is the unique achiever of the minimization in the definition of α-mutual information in (67).
    - $P_{Y\langle\alpha\rangle}$: The $\langle\alpha\rangle$-response to P_X defined in Item 26. It is the unique achiever of the minimization in the definition of Augustin–Csiszár α-mutual information in (110).
    - $Q_{X[\alpha]}$: The α-adjunct of P_X, defined in (152). The $\langle\alpha\rangle$-response to $Q_{X[\alpha]}$ is $P_{Y[\alpha]}$. Furthermore, $Q_{X[\alpha]}$ achieves the minimum in (165).
    - $Q_{X\langle\alpha\rangle}$: The $\langle\alpha\rangle$-adjunct of P_X, defined in (157). The α-response to $Q_{X\langle\alpha\rangle}$ is $P_{Y\langle\alpha\rangle}$. Furthermore, $Q_{X\langle\alpha\rangle}$ achieves the maximum in (169).

6. Maximization of $I_\alpha(X;Y)$

Just like the maximization of mutual information with respect to the input distribution yields the channel capacity (of course, subject to conditions [57]), the maximization of $I_\alpha(X;Y)$ and of $I_\alpha^{\mathsf{c}}(X;Y)$ arises in the analysis of error exponents, as we will see in Section 10. A recent in-depth treatment of the maximization of α-mutual information is given in [45]. As we see most clearly in (82) for the discrete case, when it comes to its optimization, one advantage of $I_\alpha(X;Y)$ over $I(X;Y)$ is that the input distribution does not affect the expression through its influence on the output distribution.

48. The maximization of α-mutual information is facilitated by the following result.

Theorem 14 ([45]). *Given $\alpha \in (0,1) \cup (1,\infty)$; a random transformation $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$; and, a convex set $\mathcal{P} \subset \mathcal{P}_{\mathcal{A}}$, the following are equivalent.*

(a) $P_X^* \in \mathcal{P}$ *attains the maximal α-mutual information on $\mathcal{P}$,*

$$I_\alpha(P_X^*, P_{Y|X}) = \max_{P \in \mathcal{P}} I_\alpha(P, P_{Y|X}) < \infty. \tag{177}$$

(b) *For any $P_X \in \mathcal{P}$, and any output distribution $Q_Y \in \mathcal{P}_{\mathcal{B}}$,*

$$D_\alpha(P_{Y|X} \,\|\, P_{Y[\alpha]}^* | P_X) \leqslant D_\alpha(P_{Y|X} \,\|\, P_{Y[\alpha]}^* | P_X^*) \tag{178}$$

$$\leqslant D_\alpha(P_{Y|X} \,\|\, Q_Y | P_X^*), \tag{179}$$

where $P_{Y[\alpha]}^*$ is the α-response to P_X^*.

Moreover, if $P_{Y[\alpha]}$ denotes the α-response to P_X, then

$$D_\alpha(P_{Y[\alpha]} \| P_{Y[\alpha]}^*) \leqslant I_\alpha(P_X^*, P_{Y|X}) - I_\alpha(P_X, P_{Y|X}) < \infty. \tag{180}$$

Note that, while $I_\alpha(\cdot, P_{Y|X})$ may not be maximized by a unique (or, in fact, by any) input distribution, the resulting α-response $P_{Y[\alpha]}^*$ is indeed unique. If $\mathcal{P}$ is such that none of its elements attain the maximal I_α, it is known [42,45] that the α-response to any asymptotically optimal sequence of input distributions converges to $P_{Y[\alpha]}^*$. This is the counterpart of a result by Kemperman [58] concerning mutual information.

49. The following example appears in [45].

Example 11. *Let $Y = X + N$ where $N \sim \mathcal{N}(0, \sigma_N^2)$ independent of X. Fix $\alpha \in (0,1)$ and $P > 0$. Suppose that the set, $\mathcal{P} \subset \mathcal{P}_{\mathcal{A}}$, of allowable input probability measures consists of those that satisfy the constraint*

$$\mathbb{E}\left[\exp_{\mathrm{e}}\left(-\frac{\alpha(1-\alpha)X^2}{2(\alpha^2 P + \sigma_N^2)}\right)\right] \geqslant \sqrt{\frac{\alpha^2 P + \sigma_N^2}{\alpha\, P + \sigma_N^2}}. \tag{181}$$

We can readily check that $X^* \sim \mathcal{N}(0, P)$ satisfies (181) with equality, and as we saw in Example 2, its α-response is $P_{Y[\alpha]}^* = \mathcal{N}(0, \alpha\, P + \sigma^2)$. Theorem 14 establishes that P_X^* does indeed maximize the α-mutual information among all the distributions in $\mathcal{P}$, yielding (recall Example 6)

$$\max_{P_X \in \mathcal{P}} I_\alpha(X;Y) = \frac{1}{2}\log\left(1 + \frac{\alpha P}{\sigma^2}\right). \tag{182}$$

Curiously, if, instead of $\mathcal{P}$ defined by the constraint (181), we consider the more conventional $\mathcal{P} = \{X\colon \mathbb{E}[X^2] \leqslant P\}$, then the left side of (182) is unknown at present. Numerical evidence shows that it can exceed the right side by employing non-Gaussian inputs.

50. Recalling (56) and (178) implies that if P_X^* attains the finite maximal unconstrained α-mutual information and its α-response is denoted by $P_{Y[\alpha]}^*$, then,

$$\max_X I_\alpha(X;Y) = \max_{P \in \mathcal{P}} I_\alpha(P, P_{Y|X}) = \max_{a \in \mathcal{A}} D_\alpha(P_{Y|X=a} \| P_{Y[\alpha]}^*), \tag{183}$$

which requires that $P_X^*(\mathcal{A}_\alpha^*) = 1$, with

$$\mathcal{A}_\alpha^* = \left\{x \in \mathcal{A}\colon D_\alpha(P_{Y|X=x} \| P_{Y[\alpha]}^*) = \max_{a \in \mathcal{A}} D_\alpha(P_{Y|X=a} \| P_{Y[\alpha]}^*)\right\}. \tag{184}$$

For discrete alphabets, this requires that if $x \notin \mathcal{A}_\alpha^*$, then $P_X^*(x) = 0$, which is tantamount to

$$\sum_{y \in \mathcal{B}} P_{Y|X}^\alpha(y|x)\, \mathbb{E}^{\frac{1-\alpha}{\alpha}}\left[P_{Y|X}^\alpha(y|X^*)\right] \geqslant \exp\left(\tfrac{\alpha-1}{\alpha} I_\alpha(X^*;Y^*)\right), \tag{185}$$

with equality for all $x \in \mathcal{A}$ such that $P_X^*(x) > 0$. For finite-alphabet random transformations this observation is equivalent to Theorem 5.6.5 in [9].

51. Getting slightly ahead of ourselves, we note that, in view of (128), an important consequence of Theorem 15 below, is that, as anticipated in Item 25, the unconstrained maximization of $I_\alpha(X;Y)$ for $\alpha \in (0,1)$ can be expressed in terms of the solution to an optimization problem involving only conventional mutual information and conditional relative entropy. For $\rho \geqslant 0$,

$$\rho \sup_X I_{\frac{1}{1+\rho}}(X;Y) = \sup_X \min_{Q_{Y|X}\colon \mathcal{A}\to\mathcal{B}} \Big\{ D(Q_{Y|X}\|P_{Y|X}|P_X) + \rho\, I(P_X, Q_{Y|X}) \Big\}. \tag{186}$$

7. Unconstrained Maximization of $I_\alpha^c(X;Y)$

52. In view of the fact that it is much easier to determine the α-mutual information than the order-α Augustin–Csiszár information, it would be advantageous to show that the unconstrained maximum of $I_\alpha^c(X;Y)$ equals the unconstrained maximum of $I_\alpha(X;Y)$. In the finite-alphabet setting, in which it is possible to invoke a "minisup" theorem (e.g., see Section 7.1.7 of [59]), Csiszár [32] showed this result for $\alpha > 0$. The assumption of finite output alphabets was dropped in Theorem 1 of [42], and further generalized in Theorem 3 of the same reference. As we see next, for $\alpha \in (0,1)$, it is possible to give an elementary proof without restrictions on the alphabets.

Theorem 15. *Let $\alpha \in (0,1)$. If the suprema are over $\mathcal{P}_\mathcal{A}$, the set of all probability measures defined on the input space, then*

$$\sup_X I_\alpha^c(X;Y) = \sup_X I_\alpha(X;Y). \tag{187}$$

Proof. In view of (143), $\geqslant$ holds in (187). To show $\leqslant$, we assume $\sup_X I_\alpha(X;Y) < \infty$ as, otherwise, there is nothing left to prove. The unconstrained maximization identity in (183) implies

$$\begin{aligned}
\sup_X I_\alpha(X;Y) &= \sup_{a \in \mathcal{A}} D_\alpha(P_{Y|X=a}\|P_{Y[\alpha]}^*) && (188)\\
&= \sup_{P_X \in \mathcal{P}} \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X)\|P_{Y[\alpha]}^*)\Big] && (189)\\
&\geqslant \inf_{Q \in \mathcal{Q}} \sup_{P_X \in \mathcal{P}} \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X)\|Q)\Big] && (190)\\
&\geqslant \sup_{P_X \in \mathcal{P}} \inf_{Q \in \mathcal{Q}} \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X)\|Q)\Big] && (191)\\
&= \sup_X I_\alpha^c(X;Y), && (192)
\end{aligned}$$

where $P_{Y[\alpha]}^*$ is the unique α-response to any input that achieves the maximal α-mutual information, and if there is no such input, it is the limit of the α-responses to any asymptotically optimal input sequence (Item 48). □

Furthermore, if $\{X_n\}$ is asymptotically optimal for I_α, i.e., $\lim_{n\to\infty} I_\alpha(X_n;Y_n) = \sup_X I_\alpha(X;Y)$, then $\{X_n\}$ is also asymptotically optimal for I_α^{c} because for any $\delta > 0$, we can find N, such that for all $n > N$,

$$I_\alpha(X_n;Y_n) + \delta \geqslant \sup_{a\in\mathcal{A}} D_\alpha(P_{Y|X=a}\|P^*_{Y[\alpha]}) \tag{193}$$

$$\geqslant \mathbb{E}\Big[D_\alpha(P_{Y|X}(\cdot|X_n)\|P^*_{Y[\alpha]})\Big] \tag{194}$$

$$\geqslant I_\alpha^{\mathsf{c}}(X_n;Y_n) \tag{195}$$

$$\geqslant I_\alpha(X_n;Y_n). \tag{196}$$

8. Maximization of $I_\alpha^{\mathsf{c}}(X;Y)$ Subject to Average Cost Constraints

This section is at the heart of the relevance of Rényi information measures to error exponent functions.

53. Given $\alpha \in (0,1)$, $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$, a cost function $\mathsf{b}\colon \mathcal{A} \to [0,\infty)$ and real scalar $\theta \geqslant 0$, the objective is to maximize the Augustin–Csiszár mutual information allowing only those probability measures that satisfy $\mathbb{E}[\mathsf{b}(X)] \leqslant \theta$, namely,

$$\mathbb{C}_\alpha^{\mathsf{c}}(\theta) = \sup_{\substack{P_X:\\ \mathbb{E}[\mathsf{b}(X)]\leqslant\theta}} I_\alpha^{\mathsf{c}}(P_X, P_{Y|X}). \tag{197}$$

Unfortunately, identity (187) no longer holds when the maximizations over the input probability measure are cost-constrained, and, in general, we can only claim

$$\mathbb{C}_\alpha^{\mathsf{c}}(\theta) \geqslant \sup_{\substack{P_X:\\ \mathbb{E}[\mathsf{b}(X)]\leqslant\theta}} I_\alpha(P_X, P_{Y|X}). \tag{198}$$

A conceptually simple approach to solve for $\mathbb{C}_\alpha^{\mathsf{c}}(\theta)$ is to

(a) postulate an input probability measure P_X^* that achieves the supremum in (197);
(b) solve for its $\langle\alpha\rangle$-response P_Y^* using (92);
(c) show that (P_X^*, P_Y^*) is a saddle point for the game with payoff function

$$B(P_X, Q_Y) = \int D_\alpha\Big(P_{Y|X=x}\|Q_Y\Big)\,\mathrm{d}P_X, \tag{199}$$

where $Q_Y \in \mathcal{P}_\mathcal{A}$ and P_X is chosen from the convex subset of $\mathcal{P}_\mathcal{A}$ of probability measures which satisfy $\mathbb{E}[\mathsf{b}(X)] \leqslant \theta$.

Since P_Y^* is already known, by definition, to be the $\langle\alpha\rangle$-response to P_X^*, verifying the saddle point is tantamount to showing that $B(P_X, P_Y^*)$ is maximized by P_X^* among $\{P_X \in \mathcal{P}_\mathcal{A}\colon \mathbb{E}[\mathsf{b}(X)] \leqslant \theta\}$. Theorem 1 of [43] guarantees the existence of a saddle point in the case of finite input alphabets. In addition to the fact that it is not always easy to guess the optimum input P_X^* (see e.g., Section 12), the main stumbling block is the difficulty in determining the $\langle\alpha\rangle$-response to any candidate input distribution, although sometimes this is indeed feasible as we saw in Example 7.

54. Naturally, Theorem 15 implies

$$\mathbb{C}_\alpha^{\mathsf{c}}(\theta) \leqslant \sup_X I_\alpha(X;Y). \tag{200}$$

If the unconstrained maximization of $I_\alpha^{\mathsf{c}}(\cdot, P_{Y|X})$ is achieved by an input distribution $X^\star$ that satisfies $\mathbb{E}[\mathsf{b}(X^\star)] \leqslant \theta$, then equality holds in (200), which, in turn, is equal to $I_\alpha^{\mathsf{c}}(P_X^\star, P_{Y|X})$. In that case, the average cost constraint is said to be inactive. For most cost functions and random transformations of practical interest, the cost constraint is active for all $\theta > 0$. To ascertain whether it is, we simply verify whether there exists an input achieving the right side of (200), which happens to satisfy the constraint.

If so, $\mathbb{C}_\alpha^{\mathsf{c}}(\theta)$ has been found. The same holds if we can find a sequence $\{X_n\}$ such that $\mathbb{E}[\mathsf{b}(X_n)] \leqslant \theta$ and $I_\alpha(X_n;Y_n) \to \sup_X I_\alpha(X;Y)$. Otherwise, we proceed with the method described below. Thus, henceforth, we assume that the cost constraint is active.

55. The approach proposed in this paper to solve for $\mathbb{C}_\alpha^{\mathsf{c}}(\theta)$ for $\alpha \in (0,1)$ hinges on the variational representation in (172), which allows us to sidestep having to find any $\langle\alpha\rangle$-response. Note that once we set out to maximize $I_\alpha^{\mathsf{c}}(P_X, P_{Y|X})$ over $\mathcal{P} = \{P_X \in \mathcal{P}_\mathcal{A} \colon \mathbb{E}[\mathsf{b}(X)] \leqslant \theta\}$, the allowable Q_X in the maximization in (175) range over a ξ-blow-up of $\mathcal{P}$ defined by

$$\Gamma_\xi(\mathcal{P}) = \{Q_X \in \mathcal{P}_\mathcal{A} \colon \exists P_X \in \mathcal{P}, \text{ such that } D(P_X\|Q_X) \leqslant \xi\}. \tag{201}$$

As we show in Item 56, we can accomplish such an optimization by solving an unconstrained maximization of the sum of α-mutual information and a term suitably derived from the cost function.

56. It will not be necessary to solve for (176), as our goal is to further maximize (172) over P_X subject to an average cost constraint. The Lagrangian corresponding to the constrained optimization in (197) is

$$\mathbb{L}_\alpha(\nu, P_X) = I_\alpha^{\mathsf{c}}(X;Y) - \nu\,\mathbb{E}[\mathsf{b}(X)] + \nu\,\theta, \tag{202}$$

where on the left side we have omitted, for brevity, the dependence on θ stemming from the last term on the right side. The Lagrange multiplier method (e.g., [60]) implies that if X^* achieves the supremum in (197), then there exists $\nu^* \geqslant 0$ such that for all P_X on $\mathcal{A}$ and $\nu \geqslant 0$,

$$\mathbb{L}_\alpha(\nu^*, P_X) \leqslant \mathbb{L}_\alpha(\nu^*, P_X^*) \leqslant \mathbb{L}_\alpha(\nu, P_X^*). \tag{203}$$

Note from (202) that the right inequality in (203) can only be achieved if

$$\mathbb{E}[\mathsf{b}(X^*)] = \theta, \tag{204}$$

and, consequently,

$$\mathbb{C}_\alpha^{\mathsf{c}}(\theta) = \mathbb{L}_\alpha(\nu^*, P_X^*) = \min_{\nu\geqslant 0}\max_{P_X} \mathbb{L}_\alpha(\nu, P_X) = \max_{P_X}\min_{\nu\geqslant 0} \mathbb{L}_\alpha(\nu, P_X). \tag{205}$$

The pivotal result enabling us to obtain $\mathbb{C}_\alpha^{\mathsf{c}}(\theta)$ without the need to deal with Augustin–Csiszár mutual information is the following.

Theorem 16. *Given $\alpha \in (0,1)$, $\nu \geqslant 0$, $P_{Y|X} \colon \mathcal{A} \to \mathcal{B}$, and $\mathsf{b} \colon \mathcal{A} \to [0,\infty)$, denote the function*

$$\mathbb{A}_\alpha(\nu) = \max_X \left\{ I_\alpha(X;Y) + \frac{1}{1-\alpha}\log \mathbb{E}[\exp(-(1-\alpha)\nu\,\mathsf{b}(X))] \right\}. \tag{206}$$

Then,

$$\sup_{P_X \in \mathcal{P}_\mathcal{A}} \mathbb{L}_\alpha(\nu, P_X) = \nu\,\theta + \mathbb{A}_\alpha(\nu), \tag{207}$$

and

$$\mathbb{C}_\alpha^{\mathsf{c}}(\theta) = \min_{\nu\geqslant 0}\{\nu\,\theta + \mathbb{A}_\alpha(\nu)\}. \tag{208}$$

Proof. Plugging (172) into (197) we obtain, with $X \sim P_X$, and $\hat{X} \sim Q_X$,

$$\sup_{P_X \in \mathcal{P}_{\mathcal{A}}} \mathbb{L}_\alpha(\nu, P_X) = \sup_{P_X} \{ I_\alpha^{\mathsf{c}}(X;Y) - \nu\, \mathbb{E}[\mathsf{b}(X)] + \nu\, \theta \} \tag{209}$$

$$= \sup_{P_X \in \mathcal{P}_{\mathcal{A}}} \left\{ \max_{Q_X \in \mathcal{P}_{\mathcal{A}}} \left\{ I_\alpha(Q_X, P_{Y|X}) - \frac{1}{1-\alpha} D(P_X \,\|\, Q_X) \right\} - \nu\, \mathbb{E}[\mathsf{b}(X)] + \nu\, \theta \right\} \tag{210}$$

$$= \nu\, \theta + \max_{Q_X \in \mathcal{P}_{\mathcal{A}}} \left\{ I_\alpha(Q_X, P_{Y|X}) - \frac{1}{1-\alpha} \inf_{P_X} \{ D(P_X \,\|\, Q_X) + \nu(1-\alpha) \mathbb{E}[\mathsf{b}(X)] \} \right\} \tag{211}$$

$$= \nu\, \theta + \max_{Q_X \in \mathcal{P}_{\mathcal{A}}} \left\{ I_\alpha(Q_X, P_{Y|X}) + \frac{1}{1-\alpha} \log \mathbb{E}\big[\exp\big(-\nu(1-\alpha) \mathsf{b}(\hat{X}) \big) \big] \right\} \tag{212}$$

$$= \nu\, \theta + \mathbb{A}_\alpha(\nu), \tag{213}$$

where (209) and (213) follow from (202) and (206), respectively, and (212) follows by invoking Theorem 1 with $Z \sim Q_X$ and

$$g(a) = (1-\alpha) \nu\, \mathsf{b}(a), \tag{214}$$

which is nonnegative since $\alpha \in (0,1)$ and $\nu > 0$. Finally, (208) follows from (205) and (207). □

In conclusion, we have shown that the maximization of Augustin–Csiszár mutual information of order α subject to $\mathbb{E}[\mathsf{b}(X)] \leqslant \theta$ boils down to the unconstrained maximization of a Lagrangian consisting of the sum of α-mutual information and an exponential average of the cost function. Circumventing the need to deal with $\langle \alpha \rangle$-responses and with Augustin–Csiszár mutual information of order α leads to a particularly simple optimization, as illustrated in Sections 11 and 12.

57. Theorem 16 solves for the maximal Augustin–Csiszár mutual information of order α under an average cost constraint without having to find out the input probability measure P_X^* that attains it nor its $\langle \alpha \rangle$-response P_Y^* (using the notation in Item 53). Instead, it gives the solution as

$$\mathbb{C}_\alpha^{\mathsf{c}}(\theta) = \min_{\nu \geqslant 0} \left\{ \nu\, \theta + \max_X \left\{ I_\alpha(X;Y) + \frac{1}{1-\alpha} \log \mathbb{E}[\exp(-(1-\alpha) \nu\, \mathsf{b}(X))] \right\} \right\}. \tag{215}$$

Although we are not going to invoke a minimax theorem, with the aid of Theorem 9-(b) we can see that the functional within the inner brackets is concave in P_X; Furthermore, if $V \in (0,1]$, then $\log \mathbb{E}[V^\nu]$ is easily seen to be convex in ν with the aid of the Cauchy-Schwarz inequality. Before we characterize the saddle point (ν^*, Q_X^*) of the game in (215) we note that (P_X^*, P_Y^*) can be readily obtained from (ν^*, Q_X^*).

Theorem 17. *Fix $\alpha \in (0,1)$. Let $\nu^* > 0$ denote the minimizer on the right side of (215), and Q_X^* the input probability measure that attains the maximum in (206) (or (215)) for $\nu = \nu^*$. Then,*

(a) *Q_X^* is the $\langle \alpha \rangle$-adjunct of P_X^*.*

(b) *$P_Y^* = Q_{Y[\alpha]}^*$, the α-response to Q_X^*.*

(c) *$P_X^* \ll\!\gg Q_X^*$ with*

$$\imath_{P_X^* \| Q_X^*}(a) = -(1-\alpha) \nu^*\, \mathsf{b}(a) + \tau_\alpha, \quad a \in \mathcal{A}, \tag{216}$$

where τ_α is a normalizing constant ensuring that P_X^* is a probability measure.

Proof.

(a) We had already established in Theorem 13 that the maximum on the right side of (210) is achieved by the $\langle \alpha \rangle$-adjunct of P_X. In the special case $\nu = \nu^*$, such

P_X is P_X^*. Therefore, Q_X^*, the argument that achieves the maximum in (206) for $\nu = \nu^*$, is the $\langle\alpha\rangle$-adjunct of P_X^*.

(b) According to Theorem 11, the α-response to Q_X^* is the $\langle\alpha\rangle$-response to P_X^*, which is P_Y^* by definition.

(c) For $\nu = \nu^*$, P_X^* achieves the supremum in (209) and the infimum in (211). Therefore, (216) follows from Theorem 1 with $Z \sim Q_X^*$ and $g(\cdot)$ given by (214) particularized to $\nu = \nu^*$.

□

The saddle point of (215) admits the following characterization.

Theorem 18. *If $\alpha \in (0,1)$, the saddle point (ν^*, Q_X^*) of (215) satisfies*

$$\mathbb{E}\big[\mathsf{b}(\bar{X}^*)\exp\big(-(1-\alpha)\nu^*\,\mathsf{b}(\bar{X}^*)\big)\big] = \theta\,\mathbb{E}\big[\exp\big(-(1-\alpha)\nu^*\,\mathsf{b}(\bar{X}^*)\big)\big], \quad \bar{X}^* \sim Q_X^*; \tag{217}$$

$$D_\alpha\Big(P_{Y|X=a}\,\|\,Q_{Y[\alpha]}^*\Big) = \nu^*\,\mathsf{b}(a) + c_\alpha(\nu^*), \quad a \in \mathcal{A}, \tag{218}$$

where $Q_{Y[\alpha]}^*$ is the α-response to Q_X^*, and $c_\alpha(\nu^*)$ does not depend on $a \in \mathcal{A}$. Furthermore,

$$\mathbb{A}_\alpha(\nu^*) = c_\alpha(\nu^*), \tag{219}$$

$$\mathbb{C}_\alpha^{\mathsf{c}}(\theta) = \nu^*\,\theta + c_\alpha(\nu^*). \tag{220}$$

Proof. First, we show that the scalar $\nu^* \geqslant 0$ that minimizes

$$f(\nu) = \nu\,\theta + I_\alpha(Q_X^*, P_{Y|X}) + \frac{1}{1-\alpha}\log\mathbb{E}\big[\exp\big(-(1-\alpha)\nu\,\mathsf{b}(\bar{X}^*)\big)\big] \tag{221}$$

satisfies (217). If we abbreviate $V = \exp\big(-(1-\alpha)\mathsf{b}(\bar{X}^*)\big) \in (0,1]$, then the dominated convergence theorem results in

$$\frac{\mathrm{d}}{\mathrm{d}\nu}\left\{\nu\,\theta + \frac{1}{1-\alpha}\log\mathbb{E}[V^\nu]\right\} = \theta + \frac{1}{1-\alpha}\frac{\mathbb{E}[V^\nu\log V]}{\mathbb{E}[V^\nu]}. \tag{222}$$

Therefore, (217) is equivalent to $\dot{f}(\nu^*) = 0$, which is all we need on account of the convexity of $f(\cdot)$. To show (218), notice that for all $a \in \mathcal{A}$,

$$(1-\alpha)\nu^*\,\mathsf{b}(a) - \tau_\alpha = \imath_{Q_X^*\|P_X^*}(a) \tag{223}$$

$$= (1-\alpha)D_\alpha(P_{Y|X=a}\,\|\,P_Y^*) + \upsilon_\alpha, \tag{224}$$

where (223) is (216) and (224) is (157) with $P_{Y\langle\alpha\rangle} \leftarrow P_Y^*$ in view of Theorem 17-(b). In conclusion, (218) holds with

$$c_\alpha(\nu^*) = \frac{\upsilon_\alpha + \tau_\alpha}{\alpha - 1}. \tag{225}$$

Finally, (206) implies

$$\mathbb{A}_\alpha(\nu^*) = I_\alpha(Q_X^*, P_{Y|X}) + \frac{1}{1-\alpha}\log\mathbb{E}\big[\exp\big(-(1-\alpha)\nu^*\,\mathsf{b}(\bar{X}^*)\big)\big] \tag{226}$$

$$= \frac{1}{\alpha-1}\log\mathbb{E}\Big[\exp\Big((\alpha-1)D_\alpha\Big(P_{Y|X}(\cdot|\bar{X}^*)\,\|\,P_Y^*\Big)\Big)\Big] + \frac{1}{1-\alpha}\log\mathbb{E}\big[\exp\big((\alpha-1)\nu^*\,\mathsf{b}(\bar{X}^*)\big)\big] \tag{227}$$

$$= \frac{1}{\alpha-1}\log\mathbb{E}\big[\exp\big((\alpha-1)\big(\nu^*\,\mathsf{b}(\bar{X}^*) + c_\alpha(\nu^*)\big)\big)\big] + \frac{1}{1-\alpha}\log\mathbb{E}\big[\exp\big((\alpha-1)\nu^*\,\mathsf{b}(\bar{X}^*)\big)\big] \tag{228}$$

$$= c_\alpha(\nu^*), \tag{229}$$

where (227) follows from the definition of α-mutual information and Theorem 17-(b), and (228) follows from (218). Plugging (219) into (208) results in (220). □

58. Typically, the application of Theorem 18 involves
    (a) guessing the form of the auxiliary input Q_X^* (modulo some unknown parameter),
    (b) obtaining its α-response $Q_{Y[\alpha]}^*$, and
    (c) verifying that (217) and (218) are satisfied for some specific choice of the unknown parameter.

    With the same approach, we can postulate, for every $\nu \geqslant 0$, an input distribution R_X^ν, whose α-response $R_{Y[\alpha]}^\nu$ satisfies

$$D_\alpha\Big(P_{Y|X=a}\,\|\,R_{Y[\alpha]}^\nu\Big) = \nu\,\mathsf{b}(a) + c_\alpha(\nu), \quad a \in \mathcal{A}, \tag{230}$$

where the only condition we place on $c_\alpha(\nu)$ is that it not depend on $a \in \mathcal{A}$. If this is indeed the case, then the same derivation in (226)–(229) results in

$$\mathbb{A}_\alpha(\nu) = c_\alpha(\nu), \tag{231}$$

and we determine ν^* as the solution to $\theta = -\dot{c}_\alpha(\nu^*)$, in lieu of (217). Sections 11 and 12 illustrate the effortless nature of this approach to solve for $\mathbb{A}_\alpha(\nu)$. Incidentally, (230) can be seen as the α-generalization of the condition in Problem 8.2 of [48], elaborated later in [61].

9. Gallager's E_0 Functions and the Maximal Augustin–Csiszár Mutual Information

In keeping with Gallager's setting [9], we stick to discrete alphabets throughout this section.

59. In his derivation of an achievability result for discrete memoryless channels, Gallager [8] introduced the function (1), which we repeat for convenience,

$$E_0(\rho, P_X) = -\log\sum_{y\in\mathcal{B}}\left(\sum_{x\in\mathcal{A}} P_X(x)P_{Y|X}^{\frac{1}{1+\rho}}(y|x)\right)^{1+\rho}. \tag{232}$$

Comparing (82) and (232), we obtain

$$E_0(\rho, P_X) = \rho\, I_{\frac{1}{1+\rho}}(X;Y), \tag{233}$$

which, as we mentioned in Section 1, is the observation by Csiszár in [30] that triggered the third phase in the representation of error exponents. Popularized in [9], the E_0 function was employed by Shannon, Gallager and Berlekamp [10] for $\rho \geqslant 0$

and by Arimoto [62] for $\rho \in (-1,0)$ in the derivation of converse results in data transmission, the latter of which considers rates above capacity, a region in which error probability increases with blocklength, approaching one at an exponential rate. For the achievability part, [8] showed upper bounds on the error probability involving $E_0(\rho, P_X)$ for $\rho \in [0,1]$. Therefore, for rates below capacity, the α-mutual information only enters the picture for $\alpha \in (0,1)$. One exception in which Rényi divergence of order greater than 1 plays a role at rates below capacity was found by Sason [63], where a refined achievability result is shown for binary linear codes for output symmetric channels (a case in which equiprobable P_X maximizes (233)), as a function of their Hamming weight distribution.

Although Gallager did not have the benefit of the insight provided by the Rényi information measures, he did notice certain behaviors of E_0 reminiscent of mutual information. For example, the derivative of (233) with respect to ρ, at $\rho \leftarrow 0$ is equal to $I(X;Y)$. As pointed out by Csiszár in [32], in the absence of cost constraints, Gallager's E_0 function in (232) satisfies

$$\max_{P_X} E_0(\rho, P_X) = \rho \max_{X} I_{\frac{1}{1+\rho}}(X;Y) = \rho \max_{X} I^{\mathsf{c}}_{\frac{1}{1+\rho}}(X;Y), \tag{234}$$

in view of (233) and (187).

Recall that Gallager's modified E_0 function in the case of cost constraints is

$$E_0(\rho, P_X, r, \theta) = -\log \sum_{y\in\mathcal{B}} \left(\sum_{x\in\mathcal{A}} P_X(x) \exp(r\,\mathsf{b}(x) - r\,\theta) P_{Y|X}^{\frac{1}{1+\rho}}(y|x) \right)^{1+\rho}, \tag{235}$$

which, like (232) he introduced in order to show an achievability result. Up until now, no counterpart to (234) has been found with cost constraints and (235). This is accomplished in the remainder of this section.

60. In the finite alphabet case the following result is useful to obtain a numerical solution for the functional in (206). More importantly, it is relevant to the discussion in Item 61.

Theorem 19. *In the special case of discrete alphabets, the function in* (206) *is equal to*

$$\mathbb{A}_\alpha(\nu) = \max_{G} \frac{\alpha}{\alpha-1} \log \sum_{y\in\mathcal{B}} \left(\sum_{a\in\mathcal{A}} G(a)\, P_{Y|X}^{\alpha}(y|a) \right)^{\frac{1}{\alpha}}, \tag{236}$$

where the maximization is over all $G\colon \mathcal{A} \to [0,\infty)$ *such that*

$$\sum_{a\in\mathcal{A}} G(a) \exp(-(1-\alpha)\nu \mathsf{b}(a)) = 1. \tag{237}$$

Proof. Recalling (82) we have

$$\begin{aligned} &I_\alpha(X;Y) + \frac{1}{1-\alpha} \log \mathbb{E}[\exp(-(1-\alpha)\nu\, \mathsf{b}(X))] \\ &\quad = \frac{\alpha}{\alpha-1} \log \sum_{y\in\mathcal{B}} \left(\sum_{x\in\mathcal{A}} P_X(x) P_{Y|X=x}^{\alpha}(y) \right)^{\frac{1}{\alpha}} \\ &\qquad + \frac{1}{1-\alpha} \log \mathbb{E}[\exp(-(1-\alpha)\nu\, \mathsf{b}(X))] \qquad (238) \\ &\quad = \frac{\alpha}{\alpha-1} \log \sum_{y\in\mathcal{B}} \left(\frac{\mathbb{E}\left[P_{Y|X}^{\alpha}(y|X)\right]}{\mathbb{E}[\exp(-(1-\alpha)\nu \mathsf{b}(X))]} \right)^{\frac{1}{\alpha}} \qquad (239) \\ &\quad = \frac{\alpha}{\alpha-1} \log \sum_{y\in\mathcal{B}} \left(\sum_{a\in\mathcal{A}} G(a)\, P_{Y|X}^{\alpha}(y|a) \right)^{\frac{1}{\alpha}}, \qquad (240) \end{aligned}$$

where

$$G(x) = \frac{P_X(x)}{\sum_{a\in\mathcal{A}} P_X(a)\exp(-(1-\alpha)\nu\,\mathsf{b}(a))}. \tag{241}$$

□

61. We can now proceed to close the circle between the maximization of Augustin–Csiszár mutual information subject to average cost constraints (Phase 3 in Section 1) and Gallager's approach (Phase 1 in Section 1).

Theorem 20. *In the discrete alphabet case, recalling the definitions in* (202) *and* (235) *, for* $\rho > 0$,

$$\max_{P_X} E_0(\rho, P_X, r, \theta) = \rho \max_{P_X} \mathbb{L}_{\frac{1}{1+\rho}}\left(r + \frac{r}{\rho}, P_X\right), \quad r > 0; \tag{242}$$

$$\min_{r\geqslant 0} \max_{P_X} E_0(\rho, P_X, r, \theta) = \rho\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\theta), \tag{243}$$

where the maximizations are over $\mathcal{P}_{\mathcal{A}}$.

Proof. With

$$\alpha = \frac{1}{1+\rho} \quad \text{and} \quad \nu = r\frac{1+\rho}{\rho} = \frac{r}{1-\alpha}, \tag{244}$$

the maximization of (235) with the respect to the input probability measure yields

$$\max_{P_X} E_0(\rho, P_X, r, \theta)$$

$$= \max_{P_X}\left\{(1+\rho)\,r\,\theta - \log\sum_{y\in\mathcal{B}}\left(\sum_{x\in\mathcal{A}} P_X(x)\exp(r\,\mathsf{b}(x))P_{Y|X}^{\frac{1}{1+\rho}}(y|x)\right)^{1+\rho}\right\} \tag{245}$$

$$= \rho\,\nu\,\theta + \rho\max_{P_X}\frac{\alpha}{\alpha-1}\log\sum_{y\in\mathcal{B}}\left(\sum_{x\in\mathcal{A}} P_X(x)\exp((1-\alpha)\,\nu\,\mathsf{b}(x))P_{Y|X}^{\alpha}(y|x)\right)^{\frac{1}{\alpha}} \tag{246}$$

$$= \rho\,\nu\,\theta + \rho\max_{G}\frac{\alpha}{\alpha-1}\log\sum_{y\in\mathcal{B}}\left(\sum_{x\in\mathcal{A}} G(x)P_{Y|X}^{\alpha}(y|x)\right)^{\frac{1}{\alpha}} \tag{247}$$

$$= \rho\,\nu\,\theta + \rho\,\mathbb{A}_{\alpha}(\nu) \tag{248}$$

$$= \rho\max_{P_X}\mathbb{L}_{\alpha}(\nu, P_X), \tag{249}$$

where

- the maximization on the right side of (247) is over all $G\colon \mathcal{A}\to[0,\infty)$ that satisfy (237), since that constraint is tantamount to enforcing the constraint that $P_X \in \mathcal{P}_{\mathcal{A}}$ on the left side of (247);
- (248) $\Longleftarrow$ Theorem 19;
- (249) $\Longleftarrow$ Theorem 16.

The proof of (242) is complete once (244) is invoked to substitute α and ν from the right side of (249). If we now minimize the outer sides of (245)–(249) with respect to r we obtain, using (205) and (244),

$$\min_{r\geqslant 0}\max_{P_X} E_0(\rho, P_X, r, \theta) = \rho \min_{r\geqslant 0}\max_{P_X} \mathbb{L}_\alpha\left(\frac{r}{1-\alpha}, P_X\right) \tag{250}$$

$$= \rho \min_{\nu\geqslant 0}\max_{P_X} \mathbb{L}_\alpha(\nu, P_X) \tag{251}$$

$$= \rho\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\theta). \tag{252}$$

□

In p. 329 of [9], Gallager poses the unconstrained maximization (i.e., over $P_X \in \mathcal{P}_{\mathcal{A}}$) of the Lagrangian

$$E_0(\rho, P_X, r, \theta) + \gamma \sum_{a\in\mathcal{A}} P_X(a)\mathsf{b}(a) - \gamma\,\theta. \tag{253}$$

Note the apparent discrepancy between the optimizations in (243) and (253): the latter is parametrized by r and γ (in addition to ρ and θ), while the maximization on the right side of (243) does not enforce any average cost constraint. In fact, there is no disparity since Gallager loc. cit. finds serendipitously that $\gamma = 0$ regardless of r and θ, and, therefore, just one parameter is enough.

62. The raison d'être for Augustin's introduction of I^{c}_α in [36] was his quest to view Gallager's approach with average cost constraints under the optic of Rényi information measures. Contrasting (232) and (235) and inspired by the fact that, in the absence of cost constraints, (232) satisfies a variational characterization in view of (69) and (233), Augustin [36] dealt, not with (235), but with

$$\min_{Q_Y} D_\alpha(\tilde{P}_{Y|X}\|Q_Y|P_X), \quad \text{where } \tilde{P}_{Y|X=x} = P_{Y|X=x}\exp\big(r'\mathsf{b}(x)\big).$$

Assuming finite alphabets, Augustin was able to connect this quantity with the maximal $I^{\mathsf{c}}_\alpha(X;Y)$ under cost constraints in an arcane analysis that invokes a minimax theorem. This line of work was continued in Section 5 of [43], which refers to $\min_{Q_Y} D_\alpha(\tilde{P}_{Y|X}\|Q_Y|P_X)$ as the Rényi-Gallager information. Unfortunately, since $\tilde{P}_{Y|X}$ is not a random transformation, the conditional pseudo-Rényi divergence $D_\alpha(\tilde{P}_{Y|X}\|Q_Y|P_X)$ need not satisfy the key additive decomposition in Theorem 4 so the approach of [36,43] fails to establish an identity equating the maximization of Gallager's function (235) with the maximization of Augustin–Csiszár mutual information, which is what we have accomplished through a crisp and elementary analysis.

10. Error Exponent Functions

The central objects of interest in the error exponent analysis of data transmission are the functions $E_{\mathtt{sp}}(R, P_X)$ and $E_{\mathtt{r}}(R, P_X)$ of a random transformation $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$. Reflecting the three different phases referred to in Section 1, there is no unanimity in the definition of those functions. Following [48], we adopt the standard canonical Phase 2 (Section 1.2) definitions of those functions, which are given in Items 63 and 67.

63. If $R \geqslant 0$ and $P_X \in \mathcal{P}_{\mathcal{A}}$, the sphere-packing error exponent function is (e.g., (10.19) of [48])

$$E_{\mathtt{sp}}(R, P_X) = \min_{\substack{Q_{Y|X}\colon \mathcal{A}\to\mathcal{B} \\ I(P_X, Q_{Y|X}) \leqslant R}} D(Q_{Y|X} \,\|\, P_{Y|X} \,|\, P_X). \tag{254}$$

64. As a function of $R \geqslant 0$, the basic properties of (254) for fixed $(P_X, P_{Y|X})$ are as follows.

(a) If $R \geqslant I(P_X, P_{Y|X})$, then $E_{\mathtt{sp}}(R, P_X) = 0$;
(b) If $R < I(P_X, P_{Y|X})$, then $E_{\mathtt{sp}}(R, P_X) > 0$;
(c) The infimum of the arguments for which the sphere-packing error exponent function is finite is denoted by $R_\infty(P_X)$;
(d) On the interval $R \in (R_\infty(P_X), I(P_X, P_{Y|X}))$, $E_{\mathtt{sp}}(R, P_X)$ is convex, strictly decreasing, continuous, and equal to (254) where the constraint is satisfied with equality. This implies that for R belonging to that interval, we can find $\rho_R \geqslant 0$ so that for all $r \geqslant 0$,

$$E_{\mathtt{sp}}(r, P_X) \geqslant E_{\mathtt{sp}}(R, P_X) - \rho_R\, r + \rho_R\, R. \tag{255}$$

65. In view of Theorem 8 and its definition in (254), it is not surprising that $E_{\mathtt{sp}}(R, P_X)$ is intimately related to the Augustin–Csiszár mutual information, through the following key identity.

Theorem 21.

$$E_{\mathtt{sp}}(R, P_X) = \sup_{\rho \geqslant 0} \Big\{ \rho\, I^{\mathtt{c}}_{\frac{1}{1+\rho}}(X;Y) - \rho\, R \Big\}, \quad R \geqslant 0; \tag{256}$$

$$R_\infty(P_X) = I^{\mathtt{c}}_0(X;Y). \tag{257}$$

Proof. First note that $\geqslant$ holds in (256) because from (128) we obtain, for all $\rho \geqslant 0$,

$$\rho\, I^{\mathtt{c}}_{\frac{1}{1+\rho}}(X;Y) = \min_{Q_{Y|X}} \Big\{ D(Q_{Y|X} \| P_{Y|X} | P_X) + \rho\, I(P_X, Q_{Y|X}) \Big\} \tag{258}$$

$$\leqslant \min_{\substack{Q_{Y|X}: \\ I(P_X, Q_{Y|X}) \leqslant R}} \Big\{ D(Q_{Y|X} \| P_{Y|X} | P_X) + \rho\, I(P_X, Q_{Y|X}) \Big\} \tag{259}$$

$$\leqslant E_{\mathtt{sp}}(R, P_X) + \rho\, R, \tag{260}$$

where (260) follows from the definition in (254). To show $\leqslant$ in (256) for those R such that $0 < E_{\mathtt{sp}}(R, P_X) < \infty$, Property (d) in Item 64 allows us to write

$$\min_{Q_{Y|X}} \Big\{ D(Q_{Y|X} \| P_{Y|X} | P_X) + \rho_R\, I(P_X, Q_{Y|X}) \Big\} = \min_{r \geqslant 0} \big\{ E_{\mathtt{sp}}(r, P_X) + \rho_R\, r \big\} \tag{261}$$

$$\geqslant E_{\mathtt{sp}}(R, P_X) + \rho_R\, R, \tag{262}$$

where (262) follows from (255).

To determine the region where the sphere-packing error exponent is infinite and show (257), first note that if $R < I^{\mathtt{c}}_0(X;Y) = \lim_{\alpha \downarrow 0} I^{\mathtt{c}}_\alpha(X;Y)$, then $E_{\mathtt{sp}}(R, P_X) = \infty$ because for any $\rho \geqslant 0$, the function in $\{\}$ on the right side of (256) satisfies

$$\rho\, I^{\mathtt{c}}_{\frac{1}{1+\rho}}(X;Y) - \rho\, R = \rho\, I^{\mathtt{c}}_{\frac{1}{1+\rho}}(X;Y) - \rho\, I^{\mathtt{c}}_0(X;Y) + \rho\, I^{\mathtt{c}}_0(X;Y) - \rho\, R \tag{263}$$

$$\geqslant \rho\, I^{\mathtt{c}}_0(X;Y) - \rho\, R, \tag{264}$$

where (264) follows from the monotonicity of $I^{\mathtt{c}}_\alpha(X;Y)$ in α we saw in (143). Conversely, if $I^{\mathtt{c}}_0(X;Y) < R < \infty$, there exists $\epsilon \in (0,1)$ such that $I^{\mathtt{c}}_\epsilon(X;Y) < R$, which implies that in the minimization

$$I^{\mathtt{c}}_\epsilon(X;Y) = \min_{Q_{Y|X}} \Big\{ \frac{\epsilon}{1-\epsilon} D(Q_{Y|X} \| P_{Y|X} | P_X) + I(P_X, Q_{Y|X}) \Big\} \tag{265}$$

we may restrict to those $Q_{Y|X}$ such that $I(P_X, Q_{Y|X}) \leqslant R$, and consequently, $I^{\mathtt{c}}_\epsilon(X;Y) \geqslant \frac{\epsilon}{1-\epsilon} E_{\mathtt{sp}}(R, P_X)$. Therefore, to avoid a contradiction, we must have $E_{\mathtt{sp}}(R, P_X) < \infty$.

The remaining case is $I_0^{\mathsf{c}}(X;Y) = \infty$. Again, the monotonicity of the Augustin–Csiszár mutual information implies that $I_\alpha^{\mathsf{c}}(X;Y) = \infty$ for all $\alpha > 0$. So, (128) prescribes $D(Q_{Y|X}\|P_{Y|X}|P_X) = \infty$ for any $Q_{Y|X}$ is such that $I(P_X, Q_{Y|X}) < \infty$. Therefore, $E_{\mathtt{sp}}(R, P_X) = \infty$ for all $R \geqslant 0$, as we wanted to show. □

Augustin [36] provided lower bounds on error probability for codes of type P_X as a function of $I_\alpha^{\mathsf{c}}(X;Y)$ but did not state (256); neither did Csiszár in [32] as he was interested in a non-conventional parametrization (generalized cutoff rates) of the reliability function. As pointed out in p. 5605 of [64], the ingredients for the proof of (256) were already present in the hint of Problem 23 of Section II.5 of [24]. In the discrete case, an exponential lower bound on error probability for codes with constant composition P_X is given as a function of $I^{\mathsf{c}}_{\frac{1}{1+\rho}}(P_X, P_{Y|X})$ in [44,64]. As in [64], Nakiboglu [65] gives (256) as the definition of the sphere-packing function and connects it with (254) in Lemma 3 therein, within the context of discrete input alphabets.
In the discrete case, (257) is well-known (e.g., [66]), and given by (83). As pointed out in [40], $\max_X I_0^{\mathsf{c}}(X;Y)$ is the zero-error capacity with noiseless feedback found by Shannon [67], provided there is at least a pair $(a_1, a_2) \in \mathcal{A}^2$ such that $P_{Y|X=a_1} \perp P_{Y|X=a_2}$. Otherwise, the zero-error capacity with feedback is zero.

66. The critical rate, $R_{\mathtt{c}}(P_X)$, is defined as the smallest abscissa at which the convex function $E_{\mathtt{sp}}(\cdot, P_X)$ meets its supporting line of slope -1. According to (256),

$$I^{\mathsf{c}}_{\frac{1}{2}}(X;Y) = R_{\mathtt{c}}(P_X) + E_{\mathtt{sp}}(R_{\mathtt{c}}(P_X), P_X). \tag{266}$$

67. If $R \geqslant 0$ and $P_X \in \mathcal{P}_{\mathcal{A}}$, the random-coding exponent function is (e.g., (10.15) of [48])

$$E_{\mathtt{r}}(R, P_X) = \min_{Q_{Y|X}\colon \mathcal{A}\to\mathcal{B}} \Big\{ D(Q_{Y|X}\|P_{Y|X}|P_X) + [I(P_X, Q_{Y|X}) - R]^+ \Big\}, \tag{267}$$

with $[t]^+ = \max\{0, t\}$.

68. The random-coding error exponent function is determined by the sphere-packing error exponent function through the following relation, illustrated in Figure 1.

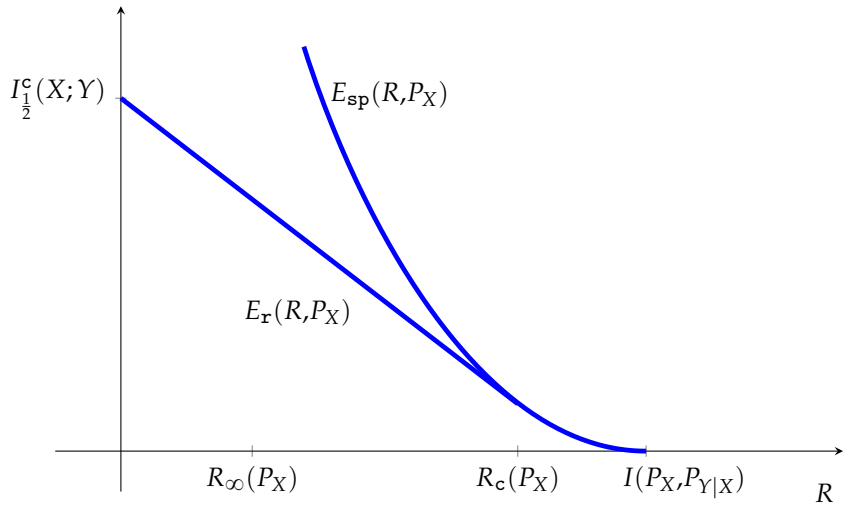

Figure 1. $E_{\mathtt{sp}}(\cdot, P_X)$ and $E_{\mathtt{r}}(\cdot, P_X)$.

Theorem 22.

$$E_{\mathtt{r}}(R,P_X) = \min_{r\geqslant R}\{E_{\mathtt{sp}}(r,P_X)+r-R\} \tag{268}$$

$$= \begin{cases} 0, & R \geqslant I(P_X,P_{Y|X}); \\ E_{\mathtt{sp}}(R,P_X), & R\in[R_{\mathtt{c}}(P_X),I(P_X,P_{Y|X})]; \\ I^{\mathtt{c}}_{\frac{1}{2}}(X;Y)-R, & R\in[0,R_{\mathtt{c}}(P_X)]. \end{cases} \tag{269}$$

$$= \sup_{\rho\in[0,1]}\left\{\rho\, I^{\mathtt{c}}_{\frac{1}{1+\rho}}(X;Y)-\rho\, R\right\}. \tag{270}$$

Proof. Identities (268) and (269) are well-known (e.g. Lemma 10.4 and Corollary 10.4 in [48]). To show (270), note that (256) expresses $E_{\mathtt{sp}}(\cdot,P_X)$ as the supremum of supporting lines parametrized by their slope $-\rho$. By definition of critical rate (for brevity, we do not show explicitly its dependence on P_X), if $R\in[R_{\mathtt{c}},I(P_X,P_{Y|X})]$, then $E_{\mathtt{sp}}(R,P_X)$ can be obtained by restricting the optimization in (256) to $\rho\in[0,1]$. In that segment of values of R, $E_{\mathtt{sp}}(R,P_X)=E_r(R,P_X)$ according to (269). Moreover, on the interval $R\in[0,R_{\mathtt{c}}]$, we have

$$\max_{\rho\in[0,1]}\left\{\rho\, I^{\mathtt{c}}_{\frac{1}{1+\rho}}(X;Y)-\rho\, R\right\} = I^{\mathtt{c}}_{\frac{1}{2}}(X;Y)-R \tag{271}$$

$$= E_{\mathtt{sp}}(R_{\mathtt{c}},P_X)+R_{\mathtt{c}}-R \tag{272}$$

$$= E_{\mathtt{r}}(R,P_X), \tag{273}$$

where we have used (266) and (269). □

The first explicit connection between $E_{\mathtt{r}}(R,P_X)$ and the Augustin–Csiszár mutual information was made by Poltyrev [35] although he used a different form for $I^{\mathtt{c}}_{\alpha}(X;Y)$, as we discussed in (29).

69. The unconstrained maximizations over the input distribution of the sphere-packing and random coding error exponent functions are denoted, respectively, by

$$E_{\mathtt{sp}}(R) = \sup_{P_X} E_{\mathtt{sp}}(R,P_X), \tag{274}$$

$$E_{\mathtt{r}}(R) = \sup_{P_X} E_{\mathtt{r}}(R,P_X). \tag{275}$$

Coding theorems [8–10,22,48] have shown that when these functions coincide they yield the reliability function (optimum speed at which the error probability vanishes with blocklength) as a function of the rate $R<\max_X I(X;Y)$. The intuition is that, for the most favorable input distribution, errors occur when the channel behaves so atypically that codes of rate R are not reliable. There are many ways in which the channel may exhibit such behavior and they are all unlikely, but the most likely among them is the one that achieves (254).
It follows from (187), (256) and (270) that (274) and (275) can be expressed as

$$E_{\mathtt{sp}}(R) = \sup_{\rho\geqslant 0}\left\{\rho\, \sup_X I_{\frac{1}{1+\rho}}(X;Y)-\rho\, R\right\}, \tag{276}$$

$$E_{\mathtt{r}}(R) = \sup_{\rho\in[0,1]}\left\{\rho\, \sup_X I_{\frac{1}{1+\rho}}(X;Y)-\rho\, R\right\}. \tag{277}$$

Therefore, we can sidestep working with the Augustin–Csiszár mutual information in the absence of cost constraints.

70. Shannon [1] showed that, operating at rates below maximal mutual information, it is possible to find codes whose error probability vanishes with blocklength; for the

converse, instead of error probability, Shannon measured reliability by the conditional entropy of the message given the channel output. That alternative reliability measure, as well as its generalization to Arimoto-Rényi conditional entropy, is also useful analyzing the average performance over code ensembles. It turns out (see e.g., [28,68]) that, below capacity, those conditional entropies also vanish exponentially fast in much the same way as error probability with bounds that are governed by $E_{\mathtt{sp}}(R)$ and $E_{\mathtt{r}}(R)$ thereby lending additional operational significance to those functions.

71. We now introduce a cost function $\mathsf{b}\colon \mathcal{A} \to [0,\infty)$ and real scalar $\theta \geqslant 0$, and reexamine the optimizations in (274) and (275) allowing only those probability measures that satisfy $\mathbb{E}[\mathsf{b}(X)] \leqslant \theta$. With a patent, but unavoidable, abuse of notation we define

$$E_{\mathtt{sp}}(R,\theta) = \sup_{\substack{P_X\colon \\ \mathbb{E}[\mathsf{b}(X)] \leqslant \theta}} E_{\mathtt{sp}}(R,P_X) \tag{278}$$

$$= \sup_{\rho \geqslant 0} \left\{ \rho \sup_{\substack{P_X\colon \\ \mathbb{E}[\mathsf{b}(X)] \leqslant \theta}} I^{\mathsf{c}}_{\frac{1}{1+\rho}}(X;Y) - \rho\, R \right\} \tag{279}$$

$$= \sup_{\rho \geqslant 0} \left\{ \rho\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\theta) - \rho\, R \right\} \tag{280}$$

$$= \sup_{\rho \geqslant 0} \left\{ -\rho\, R + \rho \min_{\nu \geqslant 0} \left\{ \nu\, \theta + \mathbb{A}_{\frac{1}{1+\rho}}(\nu) \right\} \right\} \tag{281}$$

$$= \sup_{\rho \geqslant 0} \Bigg\{ -\rho\, R + \min_{\nu \geqslant 0} \{ \rho\, \nu\, \theta + \max_{X} \left\{ \rho\, I_{\frac{1}{1+\rho}}(X;Y) + (1+\rho) \log \mathbb{E}\left[\exp\left(-\frac{\rho\,\nu}{1+\rho} \mathsf{b}(X) \right) \right] \right\} \} \Bigg\}, \tag{282}$$

where (279), (281) and (282) follow from (256), (208) and (206), respectively.

72. In parallel to (278)–(281),

$$E_{\mathtt{r}}(R,\theta) = \sup_{\substack{P_X\colon \\ \mathbb{E}[\mathsf{b}(X)] \leqslant \theta}} E_{\mathtt{r}}(R,P_X) \tag{283}$$

$$= \sup_{\rho \in [0,1]} \left\{ \rho \sup_{\substack{P_X\colon \\ \mathbb{E}[\mathsf{b}(X)] \leqslant \theta}} I^{\mathsf{c}}_{\frac{1}{1+\rho}}(X;Y) - \rho\, R \right\} \tag{284}$$

$$= \sup_{\rho \in [0,1]} \left\{ \rho\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\theta) - \rho\, R \right\}, \tag{285}$$

where (284) follows from (270). In particular, if we define the critical rate and the cutoff rate as

$$R_{\mathtt{c}} = \sup_{\substack{P_X\colon \\ \mathbb{E}[\mathsf{b}(X)] \leqslant \theta}} R_{\mathtt{c}}(P_X), \tag{286}$$

$$R_0 = \sup_{\substack{P_X\colon \\ \mathbb{E}[\mathsf{b}(X)] \leqslant \theta}} I^{\mathsf{c}}_{\frac{1}{2}}(X;Y), \tag{287}$$

respectively, then it follows from (270) that

$$E_{\mathtt{r}}(R) = R_0 - R, \quad R \in [0, R_{\mathtt{c}}]. \tag{288}$$

Summarizing, the evaluation of $E_{\mathtt{sp}}(R,\theta)$ and $E_{\mathtt{r}}(R,\theta)$ can be accomplished by the method proposed in Section 8, at the heart of which is the maximization in (206) involving α-mutual information instead of Augustin–Csiszár mutual information. In Sections 11 and 12, we illustrate the evaluation of the error exponent functions with two important additive-noise examples.

11. Additive Independent Gaussian Noise; Input Power Constraint

We illustrate the procedure in Item 58 by taking Example 6 considerably further.

73. Suppose $\mathcal{A}=\mathcal{B}=\mathbb{R}$, $\mathsf{b}(x)=x^2$, and $P_{Y|X=a}=\mathcal{N}(a,\sigma_N^2)$. We start by testing whether we can find $R_X^\nu\in\mathcal{P}_{\mathcal{A}}$ such that its α-response satisfies (230). Naturally, it makes sense to try $R_X^\nu=\mathcal{N}(0,\sigma^2)$ for some yet to be determined σ^2. As we saw in Example 6, this choice implies that its α-response is $R_{Y[\alpha]}^\nu=\mathcal{N}(0,\alpha\,\sigma^2+\sigma_N^2)$. Specializing Example 4, we obtain

$$D_\alpha\left(P_{Y|X=x}\,\|\,R_{Y[\alpha]}^\nu\right)=D_\alpha\left(\mathcal{N}\left(x,\sigma_N^2\right)\|\,\mathcal{N}\left(0,\alpha\,\sigma^2+\sigma_N^2\right)\right) \tag{289}$$

$$=\frac{1}{2}\log\left(1+\frac{\alpha\,\sigma^2}{\sigma_N^2}\right)-\frac{1}{2(1-\alpha)}\log\left(1+\frac{\alpha(1-\alpha)\sigma^2}{\alpha^2\sigma^2+\sigma_N^2}\right)+\frac{1}{2}\frac{\alpha x^2}{\alpha^2\sigma^2+\sigma_N^2}\log\mathrm{e}. \tag{290}$$

Therefore, (230) is indeed satisfied with

$$c_\alpha(\nu)=\frac{1}{2}\log\left(1+\frac{\alpha\,\sigma^2}{\sigma_N^2}\right)-\frac{1}{2(1-\alpha)}\log\left(1+\frac{\alpha(1-\alpha)\sigma^2}{\alpha^2\sigma^2+\sigma_N^2}\right), \tag{291}$$

$$\nu=\frac{1}{2}\frac{\alpha}{\alpha^2\sigma^2+\sigma_N^2}\log\mathrm{e}, \tag{292}$$

where (292) follows if we choose the variance of the auxiliary input as

$$\sigma^2=\frac{\log\mathrm{e}}{2\,\alpha\,\nu}-\frac{\sigma_N^2}{\alpha^2} \tag{293}$$

$$=\frac{\sigma_N^2}{\alpha^2}\left(\frac{\alpha}{\lambda}-1\right). \tag{294}$$

In (294) we have introduced an alternative, more convenient, parametrization for the Lagrange multiplier

$$\lambda=\frac{2\,\nu\,\sigma_N^2}{\log\mathrm{e}}\in(0,\alpha). \tag{295}$$

In conclusion, with the choice in (293), $\mathcal{N}(0,\sigma^2)$ attains the maximum in (206), and in view of (231), $\mathbb{A}_\alpha(\nu)$ is given by the right side of (291) substituting σ^2 by (293). Therefore, we have

$$\nu\,\theta+\mathbb{A}_\alpha(\nu)=\frac{\lambda}{2}\mathsf{snr}\log\mathrm{e}+c_\alpha\left(\frac{\lambda\log\mathrm{e}}{2\,\sigma_N^2}\right) \tag{296}$$

$$=\frac{\lambda}{2}\mathsf{snr}\log\mathrm{e}+\frac{1}{2}\log\left(1+\frac{1}{\lambda}-\frac{1}{\alpha}\right)-\frac{1}{2(1-\alpha)}\log(\alpha-\lambda(1-\alpha))+\frac{\log\alpha}{1-\alpha}, \tag{297}$$

where we denoted $\mathsf{snr}=\frac{\theta}{\sigma_N^2}$.

In accordance with Theorem 16 all that remains is to minimize (297) with respect to ν, or equivalently, with respect to λ. Differentiating (297) with respect to λ, the minimum is achieved at λ^* satisfying

$$\mathsf{snr} = \frac{1}{\lambda^*} \frac{\alpha - \lambda^*}{\alpha - \lambda^* + \alpha\,\lambda^*}, \tag{298}$$

whose only valid root (obtained by solving a quadratic equation) is

$$\lambda^* = \frac{1 + \alpha\,\mathsf{snr} - \alpha\,\Delta}{2\,\mathsf{snr}\,(1-\alpha)} \in (0, \alpha), \tag{299}$$

with Δ defined in (118). So, for $\alpha \in (0,1)$, (208) becomes

$$\begin{aligned} \mathbb{C}_\alpha^{\mathsf{c}}(\mathsf{snr}\,\sigma_N^2) = {} & \frac{1 + \alpha\,\mathsf{snr} - \alpha\,\Delta}{4(1-\alpha)} \log \mathrm{e} + \frac{1}{2} \log\left(1 + \frac{2\,\mathsf{snr}\,(1-\alpha)}{1 + \alpha\,\mathsf{snr} - \alpha\,\Delta} - \frac{1}{\alpha}\right) \\ & - \frac{1}{2(1-\alpha)} \log\left(\frac{\alpha\,\mathsf{snr} + \alpha\,\Delta - 1}{2\,\mathsf{snr}\,\alpha^2}\right). \end{aligned} \tag{300}$$

Letting $\alpha = \frac{1}{1+\rho}$, we obtain

$$\mathbb{C}_{\frac{1}{1+\rho}}^{\mathsf{c}}(\mathsf{snr}\,\sigma_N^2) = \frac{\mathsf{snr}}{2\,\rho}(1-\beta)\log \mathrm{e} + \frac{1}{2}\log(1 + \beta\,\mathsf{snr}) - \frac{1+\rho}{2\,\rho}\log((1+\rho)\beta), \tag{301}$$

with

$$\beta = \frac{1}{2}\left(1 - \frac{1}{\alpha\,\mathsf{snr}} + \frac{\Delta}{\mathsf{snr}}\right) = \frac{1}{2}\left(1 - \frac{1+\rho}{\mathsf{snr}} + \sqrt{\frac{4}{\mathsf{snr}} + \left(\frac{1+\rho}{\mathsf{snr}} - 1\right)^2}\right). \tag{302}$$

74. Alternatively, it is instructive to apply Theorem 18 to the current Gaussian/quadratic cost setting. Suppose we let $Q_X^* = \mathcal{N}(0, \sigma^{*2})$, where σ^{*2} is to be determined. With the aid of the formulas

$$\mathbb{E}\left[X^2\,\mathrm{e}^{-\mu X^2}\right] = \frac{\sigma^2}{\left(1 + 2\,\mu\,\sigma^2\right)^{\frac{3}{2}}}, \tag{303}$$

$$\mathbb{E}\left[\mathrm{e}^{-\mu X^2}\right] = \frac{1}{\sqrt{1 + 2\,\mu\,\sigma^2}}, \tag{304}$$

where $\mu \geqslant 0$, and $X \sim \mathcal{N}(0, \sigma^2)$, (217) becomes

$$\frac{1}{\mathsf{snr}} = \frac{\sigma_N^2}{\sigma^{*2}} + (1-\alpha)\lambda^*, \tag{305}$$

upon substituting $\sigma^2 \leftarrow \sigma^{*2}$ and

$$\mu \leftarrow \nu^* \frac{1-\alpha}{\log \mathrm{e}} = \lambda^* \frac{1-\alpha}{2\sigma_N^2}. \tag{306}$$

Likewise (218) translates into (291) and (292) with $(\nu, \sigma^2) \leftarrow (\nu^*, \sigma^{*2})$, namely,

$$c_\alpha(\nu^*) = \frac{1}{2}\log\left(1 + \frac{\alpha\,\sigma^{*2}}{\sigma_N^2}\right) - \frac{1}{2(1-\alpha)}\log\left(1 + \frac{\alpha(1-\alpha)\sigma^{*2}}{\alpha^2\sigma^{*2} + \sigma_N^2}\right), \tag{307}$$

$$\lambda^* = \frac{\alpha\sigma_N^2}{\alpha^2\sigma^{*2} + \sigma_N^2}. \tag{308}$$

Eliminating σ^{*2} from (305) by means of (308) results in (299) and the same derivation that led to (300) shows that it is equal to $\nu^*\theta + c_\alpha(\nu^*)$.

75. Applying Theorem 17, we can readily find the input distribution, P_X^*, that attains $\mathbb{C}_\alpha^{\mathsf{c}}(\theta)$ as well as its $\langle\alpha\rangle$-response P_Y^* (recall the notation in Item 53). According to Example 2, P_Y^*, the α-response to Q_X^* is Gaussian with zero mean and variance

$$\sigma_N^2 + \alpha\,\sigma^{*2} = \sigma_N^2\left(1 + \frac{1}{\lambda^*} - \frac{1}{\alpha}\right) \tag{309}$$

$$= \frac{\sigma_N^2}{2}\left(2 - \frac{1}{\alpha} + \Delta + \mathsf{snr}\right), \tag{310}$$

where (309) follows from (308) and (310) follows by using the expression for Δ in (118). Note from Example 7 that P_Y^* is nothing but the $\langle\alpha\rangle$-response to $\mathcal{N}\big(0, \mathsf{snr}\,\sigma_N^2\big)$. We can easily verify from Theorem 17 that indeed $P_X^* = \mathcal{N}\big(0, \mathsf{snr}\,\sigma_N^2\big)$ since in this case (216) becomes

$$\imath_{P_X^*\|Q_X^*}(a) = -(1-\alpha)\nu^*\,a^2 + \tau_\alpha, \tag{311}$$

which can only be satisfied by $P_X^* = \mathcal{N}\big(0, \mathsf{snr}\,\sigma_N^2\big)$ in view of (305). As an independent confirmation, we can verify, after some algebra, that the right sides of (127) and (300) are identical.

In fact, in the current Gaussian setting, we could start by postulating that the distribution that maximizes the Augustin–Csiszár mutual information under the second moment constraint does not depend on α and is given by $P_X^* = \mathcal{N}(0,\theta)$. Its $\langle\alpha\rangle$-response $P_{Y\langle\alpha\rangle}^*$ was already obtained in Example 7. Then, an alternative method to find $\mathbb{C}_\alpha^{\mathsf{c}}(\theta)$, given in Section 6.2 of [43], is to follow the approach outlined in Item 53. To validate the choice of P_X^* we must show that it maximizes $B(P_X, P_{Y\langle\alpha\rangle}^*)$ (in the notation introduced in (199)) among the subset of $\mathcal{P}_{\mathcal{A}}$ which satisfies $\mathbb{E}[X^2] \leqslant \theta$. This follows from the fact that $D_\alpha\left(P_{Y|X=x}\|P_{Y\langle\alpha\rangle}^*\right)$ is an affine function of x^2.

76. Let's now use the result in Item 73 to evaluate, with a novel parametrization, the error exponent functions for the Gaussian channel under an average power constraint.

Theorem 23. *Let $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $\mathsf{b}(x) = x^2$, and $P_{Y|X=a} = \mathcal{N}\big(a, \sigma_N^2\big)$. Then, for $\beta \in [0,1]$,*

$$E_{\mathsf{sp}}(R, \mathsf{snr}\,\sigma_N^2) = \frac{\mathsf{snr}}{2}(1-\beta)\log\mathrm{e} - \frac{1}{2}\log(1 + \mathsf{snr}\,\beta(1-\beta)), \tag{312}$$

$$R = \frac{1}{2}\log\left(1 + \frac{\beta^2}{\beta(1-\beta) + \frac{1}{\mathsf{snr}}}\right). \tag{313}$$

The critical rate and cutoff rate are, respectively,

$$R_{\mathsf{c}} = \frac{1}{2}\log\left(\frac{1}{2} + \frac{\mathsf{snr}}{4} + \frac{1}{2}\sqrt{1 + \frac{\mathsf{snr}^2}{4}}\right), \tag{314}$$

$$R_0 = \frac{1}{2}\left(1 + \frac{\mathsf{snr}}{2} - \sqrt{1 + \frac{\mathsf{snr}^2}{4}}\right)\log\mathrm{e} + \frac{1}{2}\log\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{\mathsf{snr}^2}{4}}\right). \tag{315}$$

Proof. Expression (315) for the cutoff rate follows by letting $\rho = 1$ in (301) and (302). The supremum in (281) is attained by $\rho^* \geqslant 0$ that satisfies (recall the concavity result in Theorem 9-(a))

$$R = \frac{\mathrm{d}}{\mathrm{d}\rho}\, \rho\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\mathsf{snr}\,\sigma_N^2)\bigg|_{\rho \leftarrow \rho^*} \tag{316}$$

$$= \frac{1}{2}\log\left(\mathsf{snr} + \frac{1}{\beta}\right) - \frac{1}{2}\log(1+\rho^*), \tag{317}$$

obtained after a dose of symbolic computation working with (301). In particular, letting $\rho^* = 1$, we obtain the critical rate in (314). Note that if in (302) we substitute $\rho \leftarrow \rho^*$, with ρ^* given as a function of R, snr and β by (317), we end up with an equation involving R, snr, and β. We proceed to verify that that equation is, in fact, (312). By solving a quadratic equation, we can readily check that (302) is the positive root of

$$1 + \rho = \mathsf{snr}(1-\beta) + \frac{1}{\beta}. \tag{318}$$

If we particularize (318) to $\rho \leftarrow \rho^*$, with ρ^* given by (317), namely,

$$\rho^* = -1 + \exp(-2R)\left(\mathsf{snr} + \frac{1}{\beta}\right), \tag{319}$$

we obtain

$$\exp(2R) = \frac{\mathsf{snr}\,\beta + 1}{\mathsf{snr}\,\beta(1-\beta) + 1}, \tag{320}$$

which is (313). Notice that the right side of (320) is monotonic increasing in $\beta > 0$ ranging from 1 (for $\beta = 0$) to $1 + \mathsf{snr}$ (for $\beta = 1$). Therefore, $\beta \in [0,1]$ spans the whole gamut of values of R of interest.
Assembling (281), (301) and (317), we obtain

$$E_{\mathtt{sp}}(R, \mathsf{snr}\,\sigma_N^2)$$

$$= -\rho^* R + \frac{\mathsf{snr}}{2}(1-\beta)\log \mathrm{e} + \frac{\rho^*}{2}\log(1+\beta\,\mathsf{snr}) - \frac{1+\rho^*}{2}\log((1+\rho^*)\beta) \tag{321}$$

$$= -\rho^* R + \frac{\mathsf{snr}}{2}(1-\beta)\log \mathrm{e} + \frac{\rho^*}{2}\log(1+\beta\,\mathsf{snr}) - \frac{1+\rho^*}{2}\log\beta$$
$$+ (1+\rho^*)R - \frac{1+\rho^*}{2}\log\left(\mathsf{snr} + \frac{1}{\beta}\right) \tag{322}$$

$$= R + \frac{\mathsf{snr}}{2}(1-\beta)\log \mathrm{e} - \frac{1}{2}\log(1+\beta\,\mathsf{snr}) \tag{323}$$

$$= \frac{\mathsf{snr}}{2}(1-\beta)\log \mathrm{e} - \frac{1}{2}\log(1+\mathsf{snr}\,\beta(1-\beta)), \tag{324}$$

where (324) follows by substituting (313) on the left side. □

Note that the parametric expression in (312) and (313) (shown in Figure 2) is, in fact, a closed-form expression for $E_{\mathtt{sp}}(R, \mathsf{snr}\,\sigma_N^2)$ since we can invert (313) to obtain

$$\beta = \frac{1}{2}(1-\exp(-2\,R))\left(1 + \sqrt{1 + \frac{4}{\mathsf{snr}\,(1-\exp(-2\,R))}}\right). \tag{325}$$

The random coding error exponent is

$$E_{\mathsf{r}}(R,\theta) = \begin{cases} E_{\mathsf{sp}}(R,\theta), & R \in (R_{\mathsf{c}}, \frac{1}{2}\log(1+\mathsf{snr})); \\ R_0 - R, & R \in [0, R_{\mathsf{c}}], \end{cases} \tag{326}$$

with the critical rate R_{c} and cutoff rate R_0 in (314) and (315), respectively. It can be checked that (326) coincides with the expression given by Gallager [9] (p. 340) where he optimizes (235) with respect to ρ and r, but not P_X, which he just assumes to be $P_X = \mathcal{N}(0,\theta)$. The expression for R_{c} in (314) can be found in (7.4.34) of [9]; R_0 in (314) is implicit in p. 340 of [9], and explicit in e.g., [69].

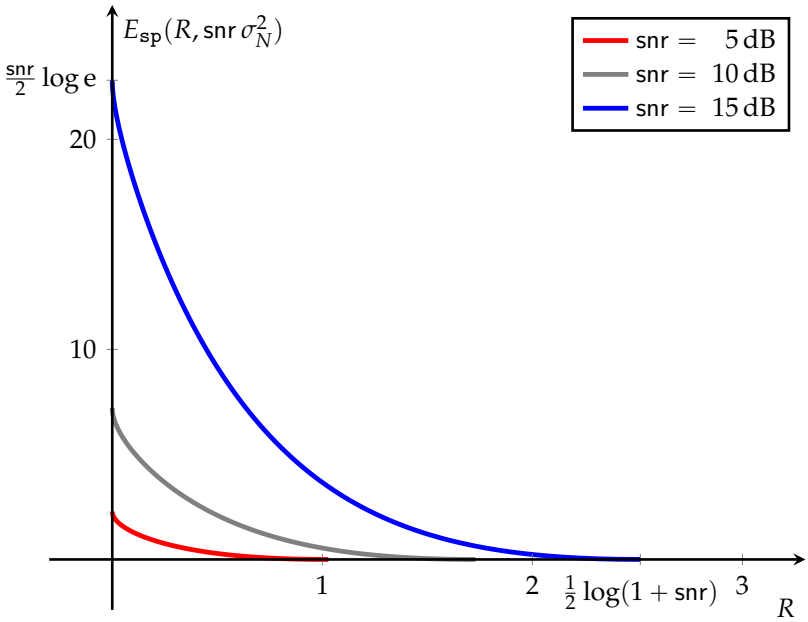

Figure 2. $E_{\mathsf{sp}}(R, \mathsf{snr}\,\sigma_N^2)$ in (312) and (313); logarithms in base 2.

77. The expression for $E_{\mathsf{sp}}(R,\theta)$ in Theorem 23 has more structure than meets the eye. The analysis in Item 73 has shown that $E_{\mathsf{sp}}(R,P_X)$ is maximized over P_X with second moment not exceeding θ by $P_X^* = \mathcal{N}(0,\theta)$ regardless of $R \in \left(0, \frac{1}{2}\log(1+\mathsf{snr})\right)$. The fact that we have found a closed-form expression for (254) when evaluated at such input probability measure and $P_{Y|X=a} = \mathcal{N}(a,\sigma_N^2)$ is indicative that the minimum therein is attained by a Gaussian random transformation $Q_{Y|X}^*$. This is indeed the case: define the random transformation

$$Q_{Y|X=a}^* = \mathcal{N}\left(\beta\, a, \sigma_1^2\right), \tag{327}$$

$$\frac{\sigma_1^2}{\sigma_N^2} = 1 + \mathsf{snr}\,\beta(1-\beta). \tag{328}$$

In comparison with the nominal random transformation $P_{Y|X=a} = \mathcal{N}(a,\sigma_N^2)$, this channel attenuates the input and contaminates it with a more powerful noise. Then,

$$I(P_X^*, Q_{Y|X}^*) = \frac{1}{2}\log\left(1 + \frac{\beta^2}{\beta(1-\beta) + \frac{1}{\mathsf{snr}}}\right) = R. \tag{329}$$

Furthermore, invoking (33), we get

$$D(Q^*_{Y|X}\|P_{Y|X}|P^*_X) = \mathbb{E}\Big[D\Big(\mathcal{N}\Big(\beta X^*, \sigma_1^2\Big)\,\|\,\mathcal{N}\Big(X^*, \sigma_N^2\Big)\Big)\Big] \tag{330}$$

$$= \frac{1}{2}\left((\beta-1)^2\mathsf{snr} + \frac{\sigma_1^2}{\sigma_N^2} - 1\right)\log \mathrm{e} - \frac{1}{2}\log\frac{\sigma_1^2}{\sigma_N^2} \tag{331}$$

$$= \frac{\mathsf{snr}}{2}(1-\beta)\log \mathrm{e} - \frac{1}{2}\log(1+\mathsf{snr}\,\beta(1-\beta)) \tag{332}$$

$$= E_{\mathsf{sp}}(R, \mathsf{snr}\,\sigma_N^2), \tag{333}$$

where (333) is (312). Therefore, $Q^*_{Y|X}$ does indeed achieve the minimum in (254) if $P_{Y|X=a} = \mathcal{N}(a, \sigma_N^2)$ and $P^*_X = \mathcal{N}(0, \theta)$. So, the most likely error mechanism is the result of atypically large noise strength and an attenuated received signal. Both effects cannot be combined into additional noise variance: there is no $\sigma^2 > 0$ such that $Q_{Y|X=a} = \mathcal{N}(a, \sigma^2)$ achieves the minimum in (254).

12. Additive Independent Exponential Noise; Input-Mean Constraint

This section finds the sphere-packing error exponent for the additive independent exponential noise channel under an input-mean constraint.

78. Suppose that $\mathcal{A} = \mathcal{B} = [0, \infty)$, $\mathsf{b}(x) = x$, and

$$Y = X + N, \tag{334}$$

where N is exponentially distributed, independent of X, and $\mathbb{E}[N] = \zeta$. Therefore $P_{Y|X=a}$ has density

$$p_{Y|X=a}(t) = \frac{1}{\zeta}\mathrm{e}^{-\frac{t-a}{\zeta}}1\{t \geqslant a\}. \tag{335}$$

It is shown in [70,71] that

$$\max_{X:\,\mathbb{E}[X]\leqslant\theta} I(X; X+N) = \log(1+\mathsf{snr}), \tag{336}$$

$$\mathsf{snr} = \frac{\theta}{\zeta}, \tag{337}$$

achieved by a mixed random variable with density

$$f^*_X(t) = \frac{\zeta}{\zeta+\theta}\,\delta(t) + \frac{\theta}{(\zeta+\theta)^2}\,\mathrm{e}^{-t/(\zeta+\theta)}1\{t>0\}. \tag{338}$$

To determine $\mathbb{C}^{\mathsf{c}}_\alpha(\mathsf{snr}\,\zeta)$, $\alpha \in (0,1)$, we invoke Theorem 18. A sensible candidate for the auxiliary input distribution Q^*_X is a mixed random variable with density

$$q^*_X(t) = \Gamma^*\,\delta(t) + (1-\Gamma^*)\frac{1}{\mu}\mathrm{e}^{-t/\mu}\,1\{t>0\}, \tag{339}$$

$$\mu = \frac{\zeta}{\alpha\,\Gamma^*}, \tag{340}$$

where $\Gamma^* \in (0,1)$ is yet to be determined. This is an attractive choice because its α-response, $Q^*_{Y[\alpha]}$, is particularly simple: exponential with mean $\alpha\,\mu = \frac{\zeta}{\Gamma^*}$, as we can

verify using Laplace transforms. Then, if Z is exponential with unit mean, with the aid of Example 5, we can write

$$D_\alpha\Big(P_{Y|X=x} \,\|\, Q^*_{Y[\alpha]}\Big) = D_\alpha(\zeta\, Z + x \,\|\, \alpha\,\mu\, Z) \tag{341}$$

$$= \frac{x}{\alpha\,\mu}\log \mathrm{e} + \log\frac{\alpha\,\mu}{\zeta} + \frac{1}{1-\alpha}\log\left(\alpha + (1-\alpha)\frac{\zeta}{\alpha\,\mu}\right) \tag{342}$$

$$= \frac{\Gamma^* x}{\zeta}\log \mathrm{e} - \log\Gamma^* + \frac{1}{1-\alpha}\log(\alpha + (1-\alpha)\Gamma^*). \tag{343}$$

So, (218) is satisfied with

$$\nu^* = \frac{\Gamma^*}{\zeta}\,\log \mathrm{e}, \tag{344}$$

$$c_\alpha(\nu^*) = \frac{1}{1-\alpha}\log(\alpha + (1-\alpha)\Gamma^*) - \log\Gamma^*. \tag{345}$$

To evaluate (217), it is useful to note that if $\gamma > -1$, then

$$\mathbb{E}\Big[Z\mathrm{e}^{-\gamma Z}\Big] = \frac{1}{(1+\gamma)^2}, \tag{346}$$

$$\mathbb{E}\Big[\mathrm{e}^{-\gamma Z}\Big] = \frac{1}{1+\gamma}. \tag{347}$$

Therefore, the left side of (217) specializes to, with $\bar{X}^* \sim Q^*_{\bar{X}}$,

$$\mathbb{E}\big[\mathsf{b}(\bar{X}^*)\exp\big(-(1-\alpha)\nu^*\,\mathsf{b}(\bar{X}^*)\big)\big] = \frac{\mu(1-\Gamma^*)}{\Big(1+\mu(1-\alpha)\frac{\nu^*}{\log \mathrm{e}}\Big)^2} \tag{348}$$

$$= \zeta\,\alpha\left(\frac{1}{\Gamma^*} - 1\right), \tag{349}$$

while the expectation on the right side of (217) is given by

$$\mathbb{E}\big[\exp\big(-(1-\alpha)\nu^*\,\mathsf{b}(\bar{X}^*)\big)\big] = \alpha + \Gamma^* - \alpha\Gamma^*. \tag{350}$$

Therefore, (217) yields

$$\mathsf{snr} = \frac{1}{\Gamma^*} - \frac{1}{\alpha + (1-\alpha)\Gamma^*} \tag{351}$$

whose solution is

$$\Gamma^* = \frac{1}{2\rho\,\mathsf{snr}}\left(\sqrt{(1+\mathsf{snr})^2 + 4\rho\,\mathsf{snr}} - 1 - \mathsf{snr}\right), \tag{352}$$

with $\rho = \frac{1-\alpha}{\alpha}$. So, finally, (220), (344) and (345) give the closed-form expression

$$\mathbb{C}^{\mathsf{c}}_\alpha(\theta) = \mathsf{snr}\,\Gamma^*\log \mathrm{e} - \log\Gamma^* + \frac{1}{1-\alpha}\log(\alpha + (1-\alpha)\Gamma^*). \tag{353}$$

As in Item 73, we can postulate an auxiliary distribution that satisfies (230) for every $\nu \geqslant 0$. This is identical to what we did in (341)–(343) except that now (344) and (345) hold for generic ν and Γ. Then, (351) is the result of solving $\theta = -\dot{c}_\alpha(\nu^*)$, which is, in fact, somewhat simpler than obtaining it through (217).

79. We proceed to get a very simple parametric expression for $E_{\mathtt{sp}}(R,\theta)$.

Theorem 24. *Let $\mathcal{A} = \mathcal{B} = [0, \infty)$, $\mathsf{b}(x) = x$, and $Y = X + N$, with N exponentially distributed, independent of X, and $\mathbb{E}[N] = \zeta$. Then, under the average cost constraint $\mathbb{E}[\mathsf{b}(X)] \leqslant \zeta\,\mathsf{snr}$,*

$$E_{\mathsf{sp}}(R, \zeta\,\mathsf{snr}) = \left(\frac{1}{\eta} - 1\right) \log \mathrm{e} + \log \eta, \tag{354}$$

$$R = \log(1 + \eta\,\mathsf{snr}), \tag{355}$$

where $\eta \in (0, 1]$.

Proof. Rewriting (353), results in

$$\rho\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\theta) = \rho\,\mathsf{snr}\,\Gamma^* \log \mathrm{e} - \rho \log \Gamma^* + (1 + \rho) \log \frac{1 + \rho\,\Gamma^*}{1 + \rho}, \tag{356}$$

which is monotonically decreasing with ρ. With $\dot{\Gamma}^* = \frac{\partial}{\partial \rho}\Gamma^*(\rho, \mathsf{snr})$, the counterpart of (317) is now

$$R = \frac{\mathrm{d}}{\mathrm{d}\rho} \rho\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\theta)\Big|_{\rho \leftarrow \rho^*} \tag{357}$$

$$= (\Gamma^* + \rho^* \dot{\Gamma}^*) \left(\mathsf{snr} - \frac{1}{\Gamma^*} + \frac{1 + \rho^*}{1 + \rho^* \Gamma^*}\right) \log \mathrm{e} + \log \frac{1 + \rho^* \Gamma^*}{\Gamma^* + \rho^*\, \Gamma^*} \tag{358}$$

$$= (\Gamma^* + \rho^* \dot{\Gamma}^*) \left(\mathsf{snr} + \frac{1}{\Gamma^*} \frac{\Gamma^* - 1}{1 + \rho^* \Gamma^*}\right) \log \mathrm{e} + \log \frac{1 + \rho^* \Gamma^*}{\Gamma^* + \rho^*\, \Gamma^*} \tag{359}$$

$$= \log \frac{1 + \rho^* \Gamma^*}{\Gamma^* + \rho^*\, \Gamma^*}, \tag{360}$$

where the drastic simplification in (360) occurs because, with the current parametrization, (351) becomes

$$1 - \Gamma^* = (1 + \rho^*\, \Gamma^*)\, \Gamma^*\, \mathsf{snr}. \tag{361}$$

Now we go ahead and express both ρ^* and Γ^* as functions of snr and R exclusively. We may rewrite (357)–(360) as

$$\rho^*\, \Gamma^* = \frac{\exp(-R) - \Gamma^*}{1 - \exp(-R)}, \tag{362}$$

which, when plugged in (361), results in

$$\Gamma^* = \frac{1}{\mathsf{snr}}(1 - \exp(-R)) < 1, \tag{363}$$

$$\rho^* = \frac{(1 + \mathsf{snr}) \exp(-R) - 1}{(1 - \exp(-R))^2} > 0, \tag{364}$$

where the inequalities in (363) and (364) follow from $R < \log(1+\mathsf{snr})$. So, in conclusion,

$$E_{\mathsf{sp}}(R,\theta) = \max_{\rho \geqslant 0} \left\{ \rho\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho}}(\theta) - \rho\, R \right\} \tag{365}$$

$$= \rho^*\, \mathbb{C}^{\mathsf{c}}_{\frac{1}{1+\rho^*}}(\theta) - \rho^*\, R \tag{366}$$

$$= \rho^*\, \mathsf{snr}\, \Gamma^*\, \log \mathrm{e} - \rho^* \log \Gamma^* + (1+\rho^*) \log \frac{1+\rho^* \Gamma^*}{1+\rho^*} - \rho^*\, R \tag{367}$$

$$= \rho^*\, \mathsf{snr}\, \Gamma^*\, \log \mathrm{e} - \rho^* \log \Gamma^* + (1+\rho^*)(R + \log \Gamma^*) - \rho^*\, R \tag{368}$$

$$= \rho^*\, \mathsf{snr}\, \Gamma^*\, \log \mathrm{e} + \log \Gamma^* + R \tag{369}$$

$$= \left(\frac{\mathsf{snr}}{\exp(R)-1} - 1 \right) \log \mathrm{e} + \log \frac{\exp(R)-1}{\mathsf{snr}} \tag{370}$$

$$= \left(\frac{1}{\eta} - 1 \right) \log \mathrm{e} + \log \eta, \tag{371}$$

where we have introduced

$$\eta = \frac{\exp(R)-1}{\mathsf{snr}} = \frac{\Gamma^*}{1-\mathsf{snr}\,\Gamma^*}. \tag{372}$$

Evidently, the left identity in (372) is the same as (355). □

The critical rate and the cutoff rate are obtained by particularizing (360) and (356) to $\rho^* = 1$ and $\rho = 1$, respectively. This yields

$$R_{\mathsf{c}} = \log \frac{1+\Gamma_1^*}{2\,\Gamma_1^*}, \tag{373}$$

$$R_0 = \mathsf{snr}\, \Gamma_1^* \log \mathrm{e} - \log(4\,\Gamma_1^*) + 2 \log(1+\Gamma_1^*), \tag{374}$$

$$\Gamma_1^* = \frac{\sqrt{(1+\mathsf{snr})^2 + 4\,\mathsf{snr}} - 1 - \mathsf{snr}}{2\,\mathsf{snr}}. \tag{375}$$

As in (326), the random coding error exponent is

$$E_{\mathsf{r}}(R, \zeta\, \mathsf{snr}) = \begin{cases} E_{\mathsf{sp}}(R, \zeta\, \mathsf{snr}), & R \in (R_{\mathsf{c}}, \log(1+\mathsf{snr})); \\ R_0 - R, & R \in [0, R_{\mathsf{c}}], \end{cases} \tag{376}$$

with the critical rate R_{c} and cutoff rate R_0 in (373) and (375), respectively. This function is shown along with $E_{\mathsf{sp}}(R, \zeta\, \mathsf{snr})$ in Figure 3 for $\mathsf{snr} = 3$.

80. In parallel to Item 77, we find the random transformation that explains the most likely mechanism to produce errors at every rate R, namely the minimizer of (254) when $P_X = P_X^*$, the maximizer of the Augustin–Csiszár mutual information of order α. In this case, P_X^* is not as trivial to guess as in Section 11, but since we already found Q_X^* in (339) with $\Gamma = \Gamma^*$, we can invoke Theorem 17 to show that the density of P_X^* achieving the maximal order-α Augustin–Csiszár mutual information is

$$p_X^*(t) = \frac{\Gamma^*}{\alpha + (1-\alpha)\Gamma^*}\, \delta(t) + \frac{1-\Gamma^*}{\alpha + (1-\alpha)\Gamma^*}\, \frac{\alpha\, \Gamma^*}{\zeta}\, \mathrm{e}^{-t\,\Gamma^*/\zeta}\, 1\{t > 0\}, \tag{377}$$

whose mean is, as it should,

$$\frac{\alpha\, \zeta}{\Gamma^*}\, \frac{1-\Gamma^*}{\alpha + (1-\alpha)\Gamma^*} = \zeta\, \mathsf{snr} = \theta. \tag{378}$$

Let Q_Y^* be exponential with mean $\theta + \kappa$, and $Q_{Y|X=a}^*$ have density

$$q_{Y|X=a}^*(t) = \frac{1}{\kappa} \mathrm{e}^{-\frac{t-a}{\kappa}} 1\{t \geqslant a\}, \tag{379}$$

with

$$\kappa = \frac{\zeta}{\eta}, \tag{380}$$

and η as defined in (372). Using Laplace transforms, we can verify that $P_X^* \to Q_{Y|X}^* \to Q_Y^*$ where P_X^* is the probability measure with density in (377). Let Z be unit-mean exponentially distributed. Writing mutual information as the difference between the output differential entropy and the noise differential entropy we get

$$I(P_X^*, Q_{Y|X}^*) = h((\theta + \kappa)Z) - h(\kappa Z) \tag{381}$$

$$= \log\left(1 + \frac{\theta}{\kappa}\right) \tag{382}$$

$$= R, \tag{383}$$

in view of (363). Furthermore, using (335) and (379),

$$D(Q_{Y|X}^* \,\|\, P_{Y|X}|P_X^*) = \log\frac{\zeta}{\kappa} + \left(\frac{\kappa}{\zeta} - 1\right)\log \mathrm{e} \tag{384}$$

$$= \log \eta + \left(\frac{1}{\eta} - 1\right)\log \mathrm{e} \tag{385}$$

$$= E_{\mathtt{sp}}(R, \zeta\,\mathsf{snr}), \tag{386}$$

where we have used (380) and (354). Therefore, we have shown that $Q_{Y|X}^*$ is indeed the minimizer of (254). In this case, the most likely mechanism for errors to happen is that the channel adds independent exponential noise with mean ζ/η, instead of the nominal mean ζ. In this respect, the behavior is reminiscent of that of the exponential timing channel for which the error exponent is dominated (at least above critical rate) by an exponential server which is slower than the nominal [72].

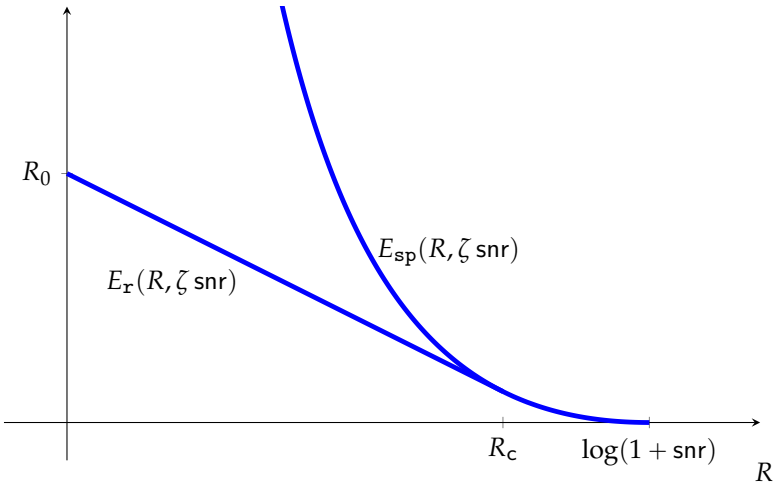

Figure 3. Error exponent functions in (354), (355) and (376).

13. Recap

81. The analysis of the fundamental limits of noisy channels in the regime of vanishing error probability with blocklength growing without bound expresses channel capacity

in terms of a basic information measure: the input–output mutual information maximized over the input distribution. In the regime of fixed nonzero error probability, the asymptotic fundamental limit is a function of not only capacity but channel dispersion [73], which is also expressible in terms of an information measure: the variance of the information density obtained with the capacity-achieving distribution. In the regime of exponentially decreasing error probability (at fixed rate below capacity) the analysis of the fundamental limits has gone through three distinct phases. No information measures were involved during the first phase and any optimization with respect to various auxiliary parameters and input distribution had to rely on standard convex optimization techniques, such as Karush-Kuhn-Tucker conditions, which not only are cumbersome to solve in this particular setting, but shed little light on the structure of the solution. The second phase firmly anchored the problem in a large deviations foundation, with the fundamental limits expressed in terms of conditional relative entropy as well as mutual information. Unfortunately, the associated maximinimization in (2) did not immediately lend itself to analytical progress. Thanks to Csiszár's realization of the relevance of Rényi's information measures to this problem, the third phase has found a way to, not only express the error exponent functions as a function of information measures, but to solve the associated optimization problems in a systematic way. While, in the absence of cost constraints, the problem reduces to finding the maximal α-mutual information, cost constraints make the problem much more challenging because of the difficulty in determining the order-α Augustin–Csiszár mutual information. Fortunately, thanks to the introduction of an auxiliary input distribution (the $\langle\alpha\rangle$-adjunct of the distribution that maximizes I_α^{c}), we have shown that α-mutual information also comes to the rescue in the maximization of the order-α Augustin–Csiszár mutual information in the presence of average cost constraints. We have also finally ended the isolation of Gallager's E_0 function with cost constraints from the representations in Phases 2 and 3. The pursuit of such a link is what motivated Augustin in 1978 to define a generalized mutual information measure. Overall, the analysis has given yet another instance of the benefits of variational representations of information measures, leading to solutions based on saddle points. However, we have steered clear of off-the-shelf minimax theorems and their associated topological constraints.

We have worked out two channels/cost constraints (additive Gaussian noise with quadratic cost, and additive exponential noise with a linear cost) that admit closed-form error-exponent functions, most easily expressed in parametric form. Furthermore, in Items 77 and 80 we have illuminated the structure of those closed-form expressions by identifying the anomalous channel behavior responsible for most errors at every given rate. In the exponential noise case, the solution is simply a noisier exponential channel, while in the Gaussian case it is the result of both a noisier Gaussian channel and an attenuated input.

These observations prompt the question of whether there might be an alternative general approach that eschews Rényi's information measures to arrive at not only the most likely anomalous channel behavior, but the error exponent functions themselves.

Funding: This research received no external funding.

Acknowledgments: The manuscript incorporates constructive suggestions by Academic Editor Igal Sason and the anonymous referees.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Recall that the relative information $\imath_{P\|Q}$ is defined only if $P \ll Q$, while $D(P\|Q) \in [0, +\infty]$ is always defined and equal to $+\infty$ if (but not only if) $P \not\ll Q$.

Lemma A1. *If $Q \ll R$ and $X \sim P \ll R$, then*

$$\mathbb{E}\Big[\imath_{P\|R}(X) - \imath_{Q\|R}(X)\Big] = D(P\,\|\,Q), \tag{A1}$$

regardless of whether the right side is finite.

Proof. If $P \ll Q \ll R$, we may invoke the chain rule (7) to decompose

$$\imath_{P\|R}(a) - \imath_{Q\|R}(a) = \imath_{P\|Q}(a). \tag{A2}$$

Then, the result follows by taking expectations of (A2) when $a \leftarrow X \sim P$.

To show that (A1) also holds when $P \not\ll Q$, i.e., that the expectation on the left side is $+\infty$, we invoke the Lebesgue decomposition theorem (e.g. p. 384 of [74]), which ensures that we can find $\alpha \in [0,1)$, $P_0 \perp Q$ and $P_1 \ll Q$, such that

$$P = \alpha\, P_1 + (1-\alpha) P_0. \tag{A3}$$

Since $P_1 \perp P_0$, we have

$$D(P_1\,\|\,P) = \log\frac{1}{\alpha}, \tag{A4}$$

$$D(P_0\,\|\,P) = \log\frac{1}{1-\alpha}. \tag{A5}$$

If $X_1 \sim P_1$, then

$$\begin{aligned}
\mathbb{E}\Big[\imath_{P\|R}(X_1) - \imath_{Q\|R}(X_1)\Big] &= \mathbb{E}\Big[\imath_{P_1\|R}(X_1) - \imath_{Q\|R}(X_1)\Big] - \mathbb{E}\Big[\imath_{P_1\|R}(X_1) - \imath_{P\|R}(X_1)\Big] && \text{(A6)}\\
&= D(P_1\,\|\,Q) - D(P_1\,\|\,P) && \text{(A7)}\\
&= D(P_1\,\|\,Q) - \log\frac{1}{\alpha}, && \text{(A8)}
\end{aligned}$$

where

- (A7) $\Longleftarrow$ (A1) with $(P,Q,R) \leftarrow (P_1,Q,R)$ and (A1) with $(P,Q,R) \leftarrow (P_1,P,R)$, which we are entitled to invoke since P_1 is dominated by both Q and R;
- (A8) $\Longleftarrow$ (A4).

Analogously, if $X_0 \sim P_0$, then

$$\begin{aligned}
\mathbb{E}\Big[\imath_{P\|R}(X_0)\Big] &= \mathbb{E}\Big[\imath_{P_0\|R}(X_0)\Big] - \mathbb{E}\Big[\imath_{P_0\|R}(X_0) - \imath_{P\|R}(X_0)\Big] && \text{(A9)}\\
&= D(P_0\,\|\,R) - D(P_0\,\|\,P) && \text{(A10)}\\
&= D(P_0\,\|\,R) - \log\frac{1}{1-\alpha}. && \text{(A11)}
\end{aligned}$$

Therefore, we are ready to conclude that

$$\begin{aligned}
&\mathbb{E}\Big[\imath_{P\|R}(X) - \imath_{Q\|R}(X)\Big]\\
&\quad= \alpha\,\mathbb{E}\Big[\imath_{P\|R}(X_1) - \imath_{Q\|R}(X_1)\Big] + (1-\alpha)\,\mathbb{E}\Big[\imath_{P\|R}(X_0) - \imath_{Q\|R}(X_0)\Big] && \text{(A12)}\\
&\quad= \alpha\, D(P_1\,\|\,Q) + (1-\alpha) D(P_0\,\|\,R) - (1-\alpha)\mathbb{E}\Big[\imath_{Q\|R}(X_0)\Big] - h(\alpha) && \text{(A13)}\\
&\quad= +\infty, && \text{(A14)}
\end{aligned}$$

where

- (A12) $\Longleftarrow$ (A3);
- (A13) $\Longleftarrow$ $h(\cdot)$ is the binary entropy function, (A8) and (A11);

- $(A14) \Longleftarrow \mathbb{E}\Big[\imath_{Q\|R}(X_0)\Big] = -\infty \Longleftarrow P_0\Big(x \in \mathcal{A}\colon \frac{\mathrm{d}Q}{\mathrm{d}R}(x) = 0\Big) = 1 \Longleftarrow P_0 \perp Q.$

□

Corollary A1. *Suppose that $Q \ll R$ and $X \sim P \ll R$. Then,*

$$\mathbb{E}\Big[\imath_{Q\|R}(X)\Big] = D(P\,\|\,R) - D(P\,\|\,Q), \tag{A15}$$

as long as at least one of the relative entropies on the right side is finite.

References

1. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
2. Rice, S.O. Communication in the Presence of Noise–Probability of Error for Two Encoding Schemes. *Bell Syst. Tech. J.* **1950**, *29*, 60–93. [CrossRef]
3. Shannon, C.E. Probability of Error for Optimal Codes in a Gaussian Channel. *Bell Syst. Tech. J.* **1959**, *38*, 611–656. [CrossRef]
4. Elias, P. Coding for Noisy Channels. *IRE Conv. Rec.* **1955**, *4*, 37–46.
5. Feinstein, A. Error Bounds in Noisy Channels without Memory. *IRE Trans. Inf. Theory* **1955**, *1*, 13–14. [CrossRef]
6. Shannon, C.E. Certain Results in Coding Theory for Noisy Channels. *Inf. Control* **1957**, *1*, 6–25. [CrossRef]
7. Fano, R.M. *Transmission of Information*; Wiley: New York, NY, USA, 1961.
8. Gallager, R.G. A Simple Derivation of the Coding Theorem and Some Applications. *IEEE Trans. Inf. Theory* **1965**, *11*, 3–18. [CrossRef]
9. Gallager, R.G. *Information Theory and Reliable Communication*; Wiley: New York, NY, USA, 1968.
10. Shannon, C.E.; Gallager, R.G.; Berlekamp, E. Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels, I. *Inf. Control* **1967**, *10*, 65–103. [CrossRef]
11. Shannon, C.E.; Gallager, R.G.; Berlekamp, E. Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels, II. *Inf. Control* **1967**, *10*, 522–552. [CrossRef]
12. Dobrushin, R.L. Asymptotic Estimates of the Error Probability for Transmission of Messages over a Discrete Memoryless Communication Channel with a Symmetric Transition Probability Matrix. *Theory Probab. Appl.* **1962**, *7*, 270–300. [CrossRef]
13. Dobrushin, R.L. Optimal Binary Codes for Low Rates of Information Transmission. *Theory Probab. Appl.* **1962**, *7*, 208–213. [CrossRef]
14. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
15. Csiszár, I.; Körner, J. Graph Decomposition: A New Key to Coding Theorems. *IEEE Trans. Inf. Theory* **1981**, *27*, 5–11. [CrossRef]
16. Barg, A.; Forney Jr., G.D. Random codes: Minimum Distances and Error Exponents. *IEEE Trans. Inf. Theory* **2002**, *48*, 2568–2573. [CrossRef]
17. Sason, I.; Shamai, S. Performance Analysis of Linear Codes under Maximum-likelihood Decoding: A Tutorial. *Found. Trends Commun. Inf. Theory* **2006**, *3*, 1–222. [CrossRef]
18. Ashikhmin, A.E.; Barg, A.; Litsyn, S.N. A New Upper Bound on the Reliability Function of the Gaussian Channel. *IEEE Trans. Inf. Theory* **2000**, *46*, 1945–1961. [CrossRef]
19. Haroutunian, E.A.; Haroutunian, M.E.; Harutyunyan, A.N. Reliability Criteria in Information Theory and in Statistical Hypothesis Testing. *Found. Trends Commun. Inf. Theory* **2007**, *4*, 97–263. [CrossRef]
20. Scarlett, J.; Peng, L.; Merhav, N.; Martinez, A.; Guillén i Fàbregas, A. Expurgated Random-coding Ensembles: Exponents, Refinements, and Connections. *IEEE Trans. Inf. Theory* **2014**, *60*, 4449–4462. [CrossRef]
21. Somekh-Baruch, A.; Scarlett, J.; Guillén i Fàbregas, A. A Recursive Cost-Constrained Construction that Attains the Expurgated Exponent. In Proceedings of the 2019 IEEE International Symposium on Information Theory, Paris, France, 7–12 July 2019; pp. 2938–2942.
22. Haroutunian, E.A. Estimates of the Exponent of the Error Probability for a Semicontinuous Memoryless Channel. *Probl. Inf. Transm.* **1968**, *4*, 29–39.
23. Blahut, R.E. Hypothesis Testing and Information Theory. *IEEE Trans. Inf. Theory* **1974**, *20*, 405–417. [CrossRef]
24. Csiszár, I.; Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*; Academic: New York, NY, USA, 1981.
25. Rényi, A. On Measures of Information and Entropy. In *Berkeley Symposium on Mathematical Statistics and Probability*; Neyman, J., Ed.; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
26. Campbell, L.L. A Coding Theorem and Rényi's Entropy. *Inf. Control* **1965**, *8*, 423–429. [CrossRef]
27. Arimoto, S. Information Measures and Capacity of Order α for Discrete Memoryless Channels. In *Topics in Information Theory*; Bolyai: Keszthely, Hungary, 1975; pp. 41–52.
28. Sason, I.; Verdú, S. Arimoto-Rényi conditional entropy and Bayesian M-ary hypothesis testing. *IEEE Trans. Inf. Theory* **2018**, *64*, 4–25. [CrossRef]
29. Fano, R.M. *Class Notes for Course 6.574: Statistical Theory of Information*; Massachusetts Institute of Technology: Cambridge, MA, USA, 1953.

30. Csiszár, I. A Class of Measures of Informativity of Observation Channels. *Period. Mat. Hung.* **1972**, *2*, 191–213. [CrossRef]
31. Sibson, R. Information Radius. *Z. Wahrscheinlichkeitstheorie Und Verw. Geb.* **1969**, *14*, 149–161. [CrossRef]
32. Csiszár, I. Generalized Cutoff Rates and Rényi's Information Measures. *IEEE Trans. Inf. Theory* **1995**, *41*, 26–34. [CrossRef]
33. Arimoto, S. Computation of Random Coding Exponent Functions. *IEEE Trans. Inf. Theory* **1976**, *22*, 665–671. [CrossRef]
34. Candan, C. Chebyshev Center Computation on Probability Simplex with α-divergence Measure. *IEEE Signal Process. Lett.* **2020**, *27*, 1515–1519. [CrossRef]
35. Poltyrev, G.S. Random Coding Bounds for Discrete Memoryless Channels. *Probl. Inf. Transm.* **1982**, *18*, 9–21.
36. Augustin, U. Noisy Channels. Ph.D. Thesis, Universität Erlangen-Nürnberg, Erlangen, Germany, 1978.
37. Tomamichel, M.; Hayashi, M. Operational Interpretation of Rényi Information Measures via Composite Hypothesis Testing against Product and Markov Distributions. *IEEE Trans. Inf. Theory* **2018**, *64*, 1064–1082. [CrossRef]
38. Polyanskiy, Y.; Verdú, S. Arimoto Channel Coding Converse and Rényi Divergence. In Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 29 September–1 October 2010; pp. 1327–1333.
39. Shayevitz, O. On Rényi Measures and Hypothesis Testing. In Proceedings of the 2011 IEEE International Symposium on Information Theory, St. Petersburg, Russia, 31 July–5 August 2011; pp. 894–898.
40. Verdú, S. α-Mutual Information. In Proceedings of the 2015 Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 1–6 February 2015.
41. Ho, S.W.; Verdú, S. Convexity/Concavity of Rényi Entropy and α-Mutual Information. In Proceedings of the 2015 IEEE International Symposium on Information Theory, Hong Kong, China, 15–19 June 2015; pp. 745–749.
42. Nakiboglu, B. The Rényi Capacity and Center. *IEEE Trans. Inf. Theory* **2019**, *65*, 841–860. [CrossRef]
43. Nakiboglu, B. The Augustin Capacity and Center. *arXiv* **2018**, arXiv:1803.07937.
44. Dalai, M. Some Remarks on Classical and Classical-Quantum Sphere Packing Bounds: Rényi vs. Kullback–Leibler. *Entropy* **2017**, *19*, 355. [CrossRef]
45. Cai, C.; Verdú, S. Conditional Rényi Divergence Saddlepoint and the Maximization of α-Mutual Information. *Entropy* **2019**, *21*, 969. [CrossRef]
46. Vázquez-Vilar, G.; Martinez, A.; Guillén i Fàbregas, A. A Derivation of the Cost-constrained Sphere-Packing Exponent. In Proceedings of the 2015 IEEE International Symposium on Information Theory, Hong Kong, China, 15–19 June 2015; pp. 929–933.
47. Wyner, A.D. Capacity and Error Exponent for the Direct Detection Photon Channel. *IEEE Trans. Inf. Theory* **1988**, *34*, 1449–1471. [CrossRef]
48. Csiszár, I.; Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2011.
49. Rényi, A. On Measures of Dependence. *Acta Math. Hung.* **1959**, *10*, 441–451. [CrossRef]
50. van Erven, T.; Harremoës, P. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
51. Csiszár, I.; Matúš, F. Information Projections Revisited. *IEEE Trans. Inf. Theory* **2003**, *49*, 1474–1490. [CrossRef]
52. Csiszár, I. Information-type Measures of Difference of Probability Distributions and Indirect Observations. *Stud. Sci. Math. Hung.* **1967**, *2*, 299–318.
53. Nakiboglu, B. The Sphere Packing Bound via Augustin's Method. *IEEE Trans. Inf. Theory* **2019**, *65*, 816–840. [CrossRef]
54. Nakiboglu, B. The Augustin Capacity and Center. *Probl. Inf. Transm.* **2019**, *55*, 299–342 [CrossRef]
55. Vázquez-Vilar, G. Error Probability Bounds for Gaussian Channels under Maximal and Average Power Constraints. *arXiv* **2019**, arXiv:1907.03163.
56. Shannon, C.E. Geometrische Deutung einiger Ergebnisse bei der Berechnung der Kanalkapazität. *Nachrichtentechnische Z.* **1957**, *10*, 1–4.
57. Verdú, S.; Han, T.S. A General Formula for Channel Capacity. *IEEE Trans. Inf. Theory* **1994**, *40*, 1147–1157. [CrossRef]
58. Kemperman, J.H.B. On the Shannon Capacity of an Arbitrary Channel. *K. Ned. Akad. Van Wet. Indag. Math.* **1974**, *77*, 101–115. [CrossRef]
59. Aubin, J.P. *Mathematical Methods of Game and Economic Theory*; North-Holland: Amsterdam, The Netherlands, 1979.
60. Luenberger, D.G. *Optimization by Vector Space Methods*; Wiley: New York, NY, USA, 1969.
61. Gastpar, M.; Rimoldi, B.; Vetterli, M. To Code, or Not to Code: Lossy Source–Channel Communication Revisited. *IEEE Trans. Inf. Theory* **2003**, *49*, 1147–1158. [CrossRef]
62. Arimoto, S. On the Converse to the Coding Theorem for Discrete Memoryless Channels. *IEEE Trans. Inf. Theory* **1973**, *19*, 357–359. [CrossRef]
63. Sason, I. On the Rényi Divergence, Joint Range of Relative Entropies, Measures and a Channel Coding Theorem. *IEEE Trans. Inf. Theory* **2016**, *62*, 23–34. [CrossRef]
64. Dalai, M.; Winter, A. Constant Compositions in the Sphere Packing Bound for Classical-quantum Channels. *IEEE Trans. Inf. Theory* **2017**, *63*, 5603–5617. [CrossRef]
65. Nakiboglu, B. The Sphere Packing Bound for Memoryless Channels. *Probl. Inf. Transm.* **2020**, *56*, 201–244. [CrossRef]
66. Dalai, M. Lower Bounds on the Probability of Error for Classical and Classical-quantum Channels. *IEEE Trans. Inf. Theory* **2013**, *59*, 8027–8056. [CrossRef]
67. Shannon, C.E. The Zero Error Capacity of a Noisy Channel. *IRE Trans. Inf. Theory* **1956**, *2*, 8–19. [CrossRef]

68. Feder, M.; Merhav, N. Relations Between Entropy and Error Probability. *IEEE Trans. Inf. Theory* **1994**, *40*, 259–266. [CrossRef]
69. Einarsson, G. Signal Design for the Amplitude-limited Gaussian Channel by Error Bound Optimization. *IEEE Trans. Commun.* **1979**, *27*, 152–158. [CrossRef]
70. Anantharam, V.; Verdú, S. Bits through Queues. *IEEE Trans. Inf. Theory* **1996**, *42*, 4–18. [CrossRef]
71. Verdú, S. The Exponential Distribution in Information Theory. *Probl. Inf. Transm.* **1996**, *32*, 86–95.
72. Arikan, E. On the Reliability Exponent of the Exponential Timing Channel. *IEEE Trans. Inf. Theory* **1996**, *48*, 1681–1689. [CrossRef]
73. Polyanskiy, Y.; Poor, H.V.; Verdú, S. Channel Coding Rate in the Finite Blocklength Regime. *IEEE Trans. Inf. Theory* **2010**, *56*, 2307–2359. [CrossRef]
74. Royden, H.L.; Fitzpatrick, P. *Real Analysis*, 4th ed.; Prentice Hall: Boston, FL, USA, 2010.