Typical Error Exponents: A Dual Domain Derivation

Giuseppe Cocco, *Senior Member, IEEE*, Albert Guillén i Fàbregas, *Fellow, IEEE*,
and Josep Font-Segura, *Member, IEEE*

Abstract— This paper shows that the probability that the error exponent of a given code randomly generated from a pairwise-independent ensemble is smaller than a lower bound on the typical random-coding exponent tends to zero as the codeword length tends to infinity. This lower bound is known to be tight for i.i.d. ensembles over the binary symmetric channel and for constant-composition codes over memoryless channels. Our results recover both as special cases and remain valid for arbitrary alphabets, arbitrary channels—for example finite-state channels with memory—, and arbitrary pairwise-independent ensembles. We specialize our results to the i.i.d., constant-composition and cost-constrained ensembles over discrete memoryless channels and to ensembles over finite-state channels.

Index Terms— Error exponents, typical random codes, typical error exponent, expurgated bound.

I. INTRODUCTION

SINCE Shannon's work [1], the random coding method remains a key tool in information theory and is applied to a wide variety of communication and compression settings. Specifically, by analysing the probability of error averaged over a certain ensemble of randomly generated codes, it is possible to show the existence of sequences of codes with a certain rate having vanishingly small probability of error.

A refinement of the random coding analysis allows to show that, for any discrete memoryless channel (DMC), the ensemble average error probability decays exponentially according to a certain error exponent, often referred to as the random-coding exponent [2], [3]. The resulting exponent is known to be tight at high rates [4], but not at low rates.

By expurgating poor codebooks from the ensemble, it is possible to show that there exist sequences of codes that attain the expurgated exponent, known to be strictly better than the random coding one at low rates [3], [5], [6]. While it is difficult to infer the structure of the resulting expurgated ensemble, it is useful to prove the existence of good low-rate codes. The expurgated exponent is known to be tight at rate that tends to zero [4]. It is also possible to show the existence of codes that attain the maximum of the expurgated and random-coding exponents [6] and to construct explicit random-coding ensembles that attain the same maximum [7].

As opposed to the exponent of the ensemble-average error probability, Barg and Forney [8] studied the average error exponent attained by the i.i.d. random-coding ensemble over the binary symmetric channel (BSC). They coined the resulting exponent the typical error exponent and showed that at low rates, it can be strictly higher than the random-coding exponent and strictly lower than the expurgated —the typical and expurgated error exponents coincide for rate zero. Nazari et al. [9] derived bounds to the typical error exponent with constant-composition ensembles for discrete-memoryless and multiple-access channels. Significant progress in the understanding of typical error exponents has been made since. Merhav has derived the typical error exponent in a number of settings, including DMCs with constant-composition codes [10], trellis codes [11], and power-constrained ensembles over the coloured Gaussian noise channel [12]. Recently, Tamir et al. [13] showed that the probability of finding a code whose exponent is lower (higher) than the typical error exponent decays exponentially (resp. double-exponentially) illustrating an asymmetry in the lower and upper tails of the distribution of the error exponent of random constant-composition codes. The concentration properties of the error exponent of randomly generated codes is studied in [14]. Tamir and Merhav [15] have shown that the typical error exponent can be achieved by a universal decoder ignorant of the channel law. Specifically, they have shown that a stochastic decoder based on the empirical mutual information between the received sequence and each codeword attains the typical error exponent. Most aforementioned derivations of the typical error exponent are done in the primal domain, i.e., they minimize a certain objective function involving information quantities such as the mutual information, entropy or the relative entropy over a constrained set of probability distributions.

In this paper, we refine a Lemma by Gallager to provide a derivation of a lower bound to the typical error exponent

for general channels and pairwise-independent random-coding ensembles. As opposed to most previous works, ours is naturally a dual-domain derivation [3], i.e., it is expressed as a maximization of an exponent function over certain non-negative real-valued parameters. This usually has the advantage that any choice of parameters provides an achievable exponent, and that it naturally extends to arbitrary alphabets. In [16] Lagrange duality was applied to the primal domain expression to obtain a dual domain expression for a lower bound on the typical error exponent for independently and identically distributed (i.i.d.) random coding and mismatched stochastic decoding. The overall expression involves an optimization over five parameters. Unlike [16], we directly derive a dual domain achievability proof of the typical error exponent. Our lower bound is valid for any channel and code ensemble with pairwise-independent codewords, thus significantly extending previous results, including [16]. We particularize our result to i.i.d., constant composition and cost-constrained ensembles over DMCs and we show that these coincide with well known expressions for the typical error exponent for DMCs, i.e., is tight in such cases. As a consequence of our derivation, we provide a rather direct dual-domain derivation of the expurgated exponent for constant composition ensembles. We also particularize our bound to the case of finite-state channels (FSC). Although it is known that reliable communication over a finite-state channel is theoretically possible at any rate below capacity [3, pp. 176–182], comparatively little research has been carried out with respect to DMC [17]. As pointed out in [18], the most studied channel model with memory has been the finite-state channel (FSC) and some of its special cases. The channel coding theorem for finite-state indecomposable channels was first proved by Blackwell, Breiman and Thomasian [19]. The random-coding exponent for FSC was studied in [20] and [21] and further developed in [3, Sec. 5.9]. In this paper, we provide an expression for the corresponding lower bound on the typical error exponent. Finally, we present an extension of our results to the case of a mismatched decoder, of which the maximum likelihood (ML) is a special case, which is valid for all channels and pairwise-independent code ensembles.

The paper is structured as follows. Section II introduces the preliminaries and the main notation used in the paper. Section III introduces the refinement of Gallager's lemma and derives the main results of the paper. Section IV particularizes the results of Section III to memoryless channels with i.i.d., constant-composition and cost-constrained code ensembles. Section V shows how to adapt the main results in Section III to finite-state channels. Proofs of some lemmas can be found in the Appendix.

II. PRELIMINARIES

We consider coding over discrete channels with conditional probability distribution $W^n(\boldsymbol{y}|\boldsymbol{x})$, being $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$ the transmitted and received sequences of length n, and $\mathcal{X}, \mathcal{Y}$ the finite channel input and output alphabets, respectively. For memoryless channels we have $W^n(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} W(y_i|x_i)$, where $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$. A code $\mathcal{C}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M_n}\}$

is a set of M_n codewords of length n. We consider a maximum-likelihood decoder that outputs message estimate $\hat{m}$ as

$$\hat{m} = \underset{m \in \{1, \ldots, M_n\}}{\arg\max} W^n(\boldsymbol{y}|\boldsymbol{x}_m). \tag{1}$$

Assuming equiprobable messages, the error probability of a fixed code $\mathcal{C}_n$ is given by

$$P_{\mathrm{e}}(\mathcal{C}_n) = \frac{1}{M_n} \sum_{m=1}^{M_n} P_{\mathrm{e}}(\mathcal{C}_n, m), \tag{2}$$

where $P_{\mathrm{e}}(\mathcal{C}_n, m)$ is the error probability conditioned to codeword $\boldsymbol{x}_m$ being transmitted. We define the finite-length error exponent of a code $\mathcal{C}_n$ as

$$E_n(\mathcal{C}_n) = -\frac{1}{n} \log P_{\mathrm{e}}(\mathcal{C}_n). \tag{3}$$

An exponent E is said to be achievable when there exists a sequence of codes $\{\mathcal{C}_n\}_{n=1}^{\infty}$ such that $\liminf_{n \to \infty} E_n(\mathcal{C}_n) \geq E$.

Lower bounds on the error exponent of codes used over discrete memoryless channels are traditionally derived using random-coding arguments [3, Sec. 5.6], [22, Ch. 10]. More specifically, let C_n be the random variable representing a code randomly generated according to some probability distribution. In code ensembles with pairwise-independent codewords, such as the i.i.d., the constant-composition or the cost-constrained ensembles later discussed in Sec. IV, the M_n codewords are generated independently with some probability distribution $Q^n(\boldsymbol{x})$. For such ensembles there exists a sequence of codes $\{\mathcal{C}_n\}_{n=1}^{\infty}$ whose limiting error exponent, given in (3) as $n \to \infty$, is at least as large as the random-coding error exponent $E_{\mathrm{r}}(R, Q)$ given by

$$E_{\mathrm{r}}(R, Q) = \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{E}[P_{\mathrm{e}}(\mathsf{C}_n)], \tag{4}$$

where $R = \lim_{n \to \infty} \frac{1}{n} \log M_n$ is the code rate and Q is the asymptotic single-letter version of Q^n. For the i.i.d., constant composition and cost-constrained ensembles over DMCs, (4) is the actual error exponent of the ensemble average error probability, and not only a lower bound. The random-coding error exponent in (4) is known to be tight at high rates [4] while it is not at low rates. Expurgation allows to show the existence of a codebook with an improved exponent, the expurgated exponent $E_{\mathrm{ex}}(R, Q)$ [3, Eq. (5.7.11)].

In contrast to the limiting exponent of the ensemble-average error probability, Barg and Forney in [8] defined the typical random-coding error exponent $E_{\mathrm{trc}}(R, Q)$ as the limiting expected error exponent of the ensemble, that is

$$E_{\mathrm{trc}}(R, Q) = \lim_{n \to \infty} -\frac{1}{n} \mathbb{E}[\log P_{\mathrm{e}}(\mathsf{C}_n)]. \tag{5}$$

The typical error exponent improves over the random-coding error exponent (4) at low rates, and is achieved by most codes in the specified ensemble, unlike the expurgated exponent, achieved by a subset of codes with unknown structure. Therefore, the typical error exponent emerges as the error exponent attained by a given ensemble for a coding length that tends to infinity. Interestingly, the expressions of the typical

and expurgated exponents are strongly connected. For the i.i.d. [8] and the constant-composition [10], [9, Lemma 3] ensembles, the typical error exponent is shown to satisfy $E_{\text{trc}}(R,Q) = E_{\text{ex}}(2R,Q) + R \le E_{\text{ex}}(R,Q)$, with equality for $R = 0$. In [10] an inequality sign is used, but this is only because the improved expurgated presented in [23, Sec. 1, point 4] is used instead of Gallager's.

III. MAIN RESULT

The main result of this paper is a consequence of the following one, a refinement of a Lemma by Gallager in [3, p. 151].

Lemma 1: Let γ_n be a sequence in n taking values in $\mathbb{R}^+$. For an arbitrary random-coding ensemble and channel and $s > 0$, it holds that

$$\mathbb{P}\left[P_{\text{e}}(\mathsf{C}_n) \ge \gamma_n^{\frac{1}{s}} \mathbb{E}[P_{\text{e}}(\mathsf{C}_n)^s]^{\frac{1}{s}}\right] \le \frac{1}{\gamma_n}. \quad (6)$$

Proof: For any $s > 0$, consider the random variable $Z = P_{\text{e}}(\mathsf{C}_n)^s$ and let $a = \gamma_n \mathbb{E}[Z]$ where γ_n is a positive real-valued sequence. Markov's inequality implies that $\mathbb{P}[Z \ge a] = \mathbb{P}[Z^{\frac{1}{s}} \ge a^{\frac{1}{s}}] \le \frac{\mathbb{E}[Z]}{a}$, obtaining (6). ∎

While Lemma 1 holds for any positive-valued sequence γ_n, by adding the constraint that γ_n be monotonically increasing such that $\lim_{n \to \infty} \gamma_n = \infty$, we obtain that the probability to randomly generate a code $\mathcal{C}_n$ such that

$$P_{\text{e}}(\mathcal{C}_n) < \gamma_n^{\frac{1}{s}} \mathbb{E}[P_{\text{e}}(\mathsf{C}_n)^s]^{\frac{1}{s}} \quad (7)$$

is larger than $1 - \frac{1}{\gamma_n}$, a quantity that tends to 1 as $n \to \infty$. The r.h.s. of (7) shows a strong connection with the typical random-coding exponent (5) as $\lim_{s \to 0} \log \mathbb{E}[P_{\text{e}}(\mathsf{C}_n)^s]^{\frac{1}{s}} = \mathbb{E}[\log P_{\text{e}}(\mathsf{C}_n)]$. However, since we assume that $\gamma_n > 1$ from a certain n, the bound (7) is tightened for $s < \infty$. This result is subsumed in Theorems 1 and 2, valid for channels with arbitrary alphabets or memory and pairwise-independent code ensembles.

For compactness of notation, in the following we will be using the quantity:

$$R_n \triangleq \frac{1}{n} \log M_n, \quad (8)$$

where $\lim_{n \to \infty} R_n = R$.

Before moving to Theorem 1 we define the following quantities.

Definition 1: We define

$$E_{\text{trc,x}}(R_n, Q^n) = E_{\text{ex}}^n(2R_n, Q^n) + R_n - \delta_n, \quad (9)$$

where

$$E_{\text{ex}}^n(R_n, Q^n) = E_{\text{x}}^n(\hat{\lambda}_n, Q^n) - \hat{\lambda}_n R_n \quad (10)$$

is the multi-letter version of the expurgated exponent,

$$E_{\text{x}}^n(\lambda, Q^n) = -\frac{1}{n} \log\left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') Z_n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}}\right)^{\lambda}, \quad (11)$$

$Z_n(\boldsymbol{x}, \boldsymbol{x}') = \sum_{\boldsymbol{y}} \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}')}$ is the Bhattacharyya coefficient between $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n$,

$$\hat{\lambda}_n = \arg\max_{\lambda \ge 1}\{E_{\text{x}}^n(\lambda, Q^n) - \lambda 2R_n\} \quad (12)$$

is the bound parameter that yields the highest exponent, while

$$\delta_n = \frac{\hat{\lambda}_n}{n} \log \gamma_n, \quad (13)$$

and γ_n is a positive-defined sequence.

Definition 2: Let

$$E_{\text{trc,r}}(R_n, Q^n) = E_{\text{r}}^n(R_n, Q^n) - \iota_n, \quad (14)$$

where

$$E_{\text{r}}^n(R_n, Q^n) = E_0(\hat{\rho}_n, Q^n) - \hat{\rho}_n R_n \quad (15)$$

is the multi-letter version of Gallager's random-coding exponent in [3, Eq. (5.6.16)],

$$E_0^n(\rho, Q^n) = -\frac{1}{n} \log\left(\sum_{\boldsymbol{y}}\left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}}\right)^{1+\rho}\right), \quad (16)$$

$\hat{\rho}_n$ is the optimal bound parameter

$$\hat{\rho}_n = \arg\max_{0 \le \rho \le 1}\{E_0(\rho, Q^n) - \rho R_n\}, \quad (17)$$

and

$$\iota_n = \frac{1}{n} \log \gamma_n, \quad (18)$$

where γ_n is a positive-defined sequence.

Theorem 1: For a channel W^n and a pairwise-independent ensemble with codeword distribution Q^n, it holds that

$$\mathbb{P}\left[E_n(\mathsf{C}_n) > E_{\text{trc,lb}}(R_n, Q^n)\right] \ge 1 - \frac{1}{\gamma_n}, \quad (19)$$

where γ_n is a positive real-valued sequence and

$$E_{\text{trc,lb}}(R_n, Q^n) = \max\left\{E_{\text{trc,x}}(R_n, Q^n), E_{\text{trc,r}}(R_n, Q^n)\right\}. \quad (20)$$

Proof: We start deriving an upper bound on the average tilted error probability.

For a given code $\mathcal{C}_n$ with equiprobable messages, we have:

$$P_{\text{e}}(\mathcal{C}_n) = \frac{1}{M_n} \sum_{m=1}^{M_n} P_{\text{e}}(\mathcal{C}_n, m) \quad (21)$$

$$\le \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x}_m)^{\frac{1}{1+\rho}} \cdot$$

$$\cdot \left(\sum_{m' \ne m} W^n(\boldsymbol{y}|\boldsymbol{x}_{m'})^{\frac{1}{1+\rho}}\right)^{\rho}, \quad (22)$$

where (22), valid for $\rho \ge 0$, is obtained from Gallager's bound to a given code [24]. Raising (22) to the power of s with range in $(0, 1]$, we have that:

$$P_{\text{e}}(\mathcal{C}_n)^s \le \left(\frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x}_m)^{\frac{1}{1+\rho}} \cdot\right.$$

$$\cdot \left(\sum_{m' \neq m} W^n(\boldsymbol{y}|\boldsymbol{x}_{m'})^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s \quad (23)$$

$$\leq \sum_{m=1}^{M_n} \left(\frac{1}{M_n} \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x}_m)^{\frac{1}{1+\rho}} \cdot \right.$$
$$\left. \cdot \left(\sum_{m' \neq m} W^n(\boldsymbol{y}|\boldsymbol{x}_{m'})^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s, \quad (24)$$

where (24) follows from the fact that $(\sum a_i)^s \leq \sum (a_i)^s$ for $s \in (0,1]$ [3, Ch. 5]. Taking the ensemble average of (24) we obtain:

$$\mathbb{E}[P_{\mathrm{e}}(\mathsf{C}_n)^s] \leq \mathbb{E}\left[\sum_{m=1}^{M_n} \left(\frac{1}{M_n} \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{X}_m)^{\frac{1}{1+\rho}} \cdot \right. \right.$$
$$\left. \left. \cdot \left(\sum_{m' \neq m} W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s \right] \quad (25)$$

$$= \sum_{m=1}^{M_n} \mathbb{E}\left[\left(\frac{1}{M_n} \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{X}_m)^{\frac{1}{1+\rho}} \cdot \right. \right.$$
$$\left. \left. \cdot \left(\sum_{m'=1}^{M_n} W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s \right] \quad (26)$$

$$= M_n \mathbb{E}\left[\left(\frac{1}{M_n} \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{X})^{\frac{1}{1+\rho}} \cdot \right. \right.$$
$$\left. \left. \cdot \left(\sum_{m'=1}^{M_n} W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s \right], \quad (27)$$

where (27) is because m is a dummy variable and, after averaging, it leads to M_n equal terms. Rising (27) to $\frac{1}{s}$, we obtain the following bound on the tilted average probability of error:

$$\mathbb{E}[P_{\mathrm{e}}(\mathsf{C}_n)^s]^{\frac{1}{s}} \leq \frac{1}{M_n} M_n^{\frac{1}{s}} \left(\mathbb{E}\left[\left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{X})^{\frac{1}{1+\rho}} \cdot \right. \right. \right.$$
$$\left. \left. \left. \cdot \left(\sum_{m'=1}^{M_n} W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s \right] \right)^{\frac{1}{s}}. \quad (28)$$

Now we proceed to upper bound the right-hand side of (7) in two different ways starting from (28). We show that one of the two bounds is $E_{\mathrm{ex}}^n(2R_n, Q^n) + R_n$, where $E_{\mathrm{ex}}^n(R_n, Q^n)$ is Gallager's expurgated exponent, while the other is the random-coding exponent. In Remark 3 after Theorem 2 we show that, under certain conditions, taking the maximum of the two and optimizing over the input distribution we recover the typical random coding (TRC) exponent for the whole range of rates below capacity.

a) Derivation of $E_{\mathrm{trc,x}}(R_n, Q^n)$: Restricting ρ to $[0,1]$ in (28) while keeping $s \in (0,1]$ we have the following set of steps

$$\mathbb{E}[P_{\mathrm{e}}(\mathsf{C}_n)^s]^{\frac{1}{s}} \leq \frac{1}{M_n} M_n^{\frac{1}{s}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \mathbb{E}\left[\left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right. \right.$$
$$\left. \left. \left. \cdot \left(\sum_{m'=1}^{M_n} W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s \middle| \boldsymbol{X} = \boldsymbol{x} \right] \right)^{\frac{1}{s}} \quad (29)$$

$$\leq \frac{1}{M_n} M_n^{\frac{1}{s}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \mathbb{E}\left[\left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right. \right.$$
$$\left. \left. \left. \cdot \sum_{m'=1}^{M_n} W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{\rho}{1+\rho}} \right)^s \middle| \boldsymbol{X} = \boldsymbol{x} \right] \right)^{\frac{1}{s}} \quad (30)$$

$$\leq \frac{1}{M_n} M_n^{\frac{1}{s}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \mathbb{E}\left[\sum_{m'=1}^{M_n} \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right. \right.$$
$$\left. \left. \left. \cdot W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{\rho}{1+\rho}} \right)^s \middle| \boldsymbol{X} = \boldsymbol{x} \right] \right)^{\frac{1}{s}} \quad (31)$$

$$= \frac{1}{M_n} M_n^{\lambda} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sum_{m'=1}^{M_n} \mathbb{E}\left[\left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{1-t} \cdot \right. \right. \right.$$
$$\left. \left. \left. \cdot W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{t} \right)^{\frac{1}{\lambda}} \middle| \boldsymbol{X} = \boldsymbol{x} \right] \right)^{\lambda} \quad (32)$$

$$= \frac{1}{M_n} (M_n^2)^{\lambda} \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}'|\boldsymbol{x}) \cdot \right.$$
$$\left. \cdot \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{1-t} W^n(\boldsymbol{y}|\boldsymbol{x}')^{t} \right)^{\frac{1}{\lambda}} \right)^{\lambda} \quad (33)$$

$$= \frac{1}{M_n} (M_n^2)^{\lambda} \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \cdot \right.$$
$$\left. \cdot \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{1-t} W^n(\boldsymbol{y}|\boldsymbol{x}')^{t} \right)^{\frac{1}{\lambda}} \right)^{\lambda} \quad (34)$$

$$\leq \frac{1}{M_n} (M_n^2)^{\lambda} \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \cdot \right.$$
$$\left. \cdot \left(\sum_{\boldsymbol{y}} \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}')} \right)^{\frac{1}{\lambda}} \right)^{\lambda} \quad (35)$$

where in (29) we conditioned the expectation to a given $\boldsymbol{x}$ and averaged out over all possible $\boldsymbol{x}$, (30) follows from the inequality $(\sum_i a_i)^{\rho} \leq \sum_i (a_i)^{\rho}$ for $0 \leq \rho \leq 1$, (31) follows from the same inequality considered with respect to s instead of ρ, equality (32) comes from renaming the variables $s \to 1/\lambda$, $\rho \to t/(1-t)$ (to make the expression aesthetically similar to Gallager's expurgated), changing their ranges to $\lambda \in [1, \infty)$, $t \in [0, 1/2]$ and from the linearity of the expectation, while (33) is because m' is a dummy variable of summation leading to M_n identical terms and from the fact that we consider pairwise-independent codewords, which implies $Q^n(\boldsymbol{x}'|\boldsymbol{x}) = Q^n(\boldsymbol{x}')$, where $Q^n(\boldsymbol{x}'|\boldsymbol{x})$ is the probability to generate codeword $\boldsymbol{x}'$ given that codeword $\boldsymbol{x}$ has been already generated. Now we note that (33) is convex in t and symmetric with respect to $t = 1/2$. Thus, (33) is minimized with respect to t in $t = 1/2$, leading to expression (35). Taking the negative normalized logarithm of (35), we obtain for $s \in (0,1]$ and $\lambda \in [1, \infty)$ that

$$-\frac{1}{n} \log \mathbb{E}[P_{\mathrm{e}}(\mathsf{C}_n)^s]^{\frac{1}{s}} \geq$$
$$R_n - \lambda 2 R_n - \frac{\lambda}{n} \log\left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') Z_n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}} \right). \quad (36)$$

Including the γ_n in (36) and optimizing both sides in the respective variables, we have

$$\max_{s \in (0,1]} \left\{ -\frac{1}{s} \frac{1}{n} \log \gamma_n - \frac{1}{n} \log \mathbb{E}[P_{\mathrm{e}}(\mathsf{C}_n)^s]^{\frac{1}{s}} \right\} \geq$$

$$\max_{\lambda \in [1,\infty)} \left\{ -\frac{\lambda}{n} \log \gamma_n + R_n - \lambda 2 R_n + \right.$$
$$\left. -\frac{\lambda}{n} \log \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') Z_n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}} \right) \right\}.$$
(37)

b) Derivation of $E_{\mathrm{trc,r}}(R_n, Q^n)$: In the following we derive the random-coding exponent starting from (27). Restricting ρ to $[0,1]$ in (28) while keeping $s \in (0,1]$ we have now that

$$\mathbb{E}[P_{\mathrm{e}}(\mathsf{C}_n)^s]^{\frac{1}{s}}$$

$$\leq \frac{1}{M_n} M_n^{\frac{1}{s}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \mathbb{E}\left[\left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right. \right.$$
$$\left. \left. \left. \cdot \left(\sum_{m'=1}^{M_n} W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s \middle| \boldsymbol{X} = \boldsymbol{x} \right] \right)^{\frac{1}{s}}$$
(38)

$$\leq \frac{1}{M_n} M_n^{\frac{1}{s}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right.$$
$$\left. \left. \cdot \mathbb{E}\left[\left(\sum_{m'=1}^{M_n} W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{1}{1+\rho}} \right)^{\rho} \middle| \boldsymbol{X} = \boldsymbol{x} \right] \right)^s \right)^{\frac{1}{s}}$$
(39)

$$\leq \frac{1}{M_n} M_n^{\frac{1}{s}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right.$$
$$\left. \left. \cdot \left(\sum_{m'=1}^{M_n} \mathbb{E}\left[W^n(\boldsymbol{y}|\boldsymbol{X}_{m'})^{\frac{1}{1+\rho}} \middle| \boldsymbol{X} = \boldsymbol{x} \right] \right)^{\rho} \right)^s \right)^{\frac{1}{s}}$$
(40)

$$= \frac{1}{M_n} M_n^{\frac{1}{s}} M_n^{\rho} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right.$$
$$\left. \left. \cdot \left(\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}'|\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}')^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s \right)^{\frac{1}{s}}$$
(41)

$$= \frac{1}{M_n} M_n^{\frac{1}{s}} M_n^{\rho} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right.$$
$$\left. \left. \cdot \left(\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}') W^n(\boldsymbol{y}|\boldsymbol{x}')^{\frac{1}{1+\rho}} \right)^{\rho} \right)^s \right)^{\frac{1}{s}}$$
(42)

$$= \frac{1}{M_n} M_n^{t} M_n^{\rho} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right.$$
$$\left. \left. \cdot \left(\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}') W^n(\boldsymbol{y}|\boldsymbol{x}')^{\frac{1}{1+\rho}} \right)^{\rho} \right)^{\frac{1}{t}} \right)^{t}$$
(43)

$$\leq \frac{1}{M_n} M_n^{t} M_n^{\rho} \left(\left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right. \right.$$
$$\left. \left. \cdot \left(\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}') W^n(\boldsymbol{y}|\boldsymbol{x}')^{\frac{1}{1+\rho}} \right)^{\rho} \right)^{\frac{1}{t}} \right)^{t}$$
(44)

$$\leq \frac{1}{M_n} M_n^{t} M_n^{\rho} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \cdot \right.$$

$$\left. \cdot \left(\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}') W^n(\boldsymbol{y}|\boldsymbol{x}')^{\frac{1}{1+\rho}} \right)^{\rho} \right)$$
(45)

$$\leq \frac{1}{M_n} M_n^{t} M_n^{\rho} \left(\sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}')^{\frac{1}{1+\rho}} \right)^{1+\rho} \right),$$
(46)

where (39) follows from Jensen's inequality and the concavity of x^s for $s \in (0,1]$, (40) follows from Jensen's inequality and the concavity of x^{ρ} for $\rho \in [0,1]$, (41) comes from the fact that m' is a dummy variable once the average over $\boldsymbol{x}'$ is taken leading to $((M_n^{\rho})^s)^{1/s} = M_n^{\rho}$ equal terms, (42) is because of pairwise independence of the codewords, in (43) we applied the change of variable $s \to 1/t$, $t \in [1, \infty)$, while (44) is because of Jensen's inequality applied to the function $x^{\frac{1}{t}}$ which is concave for $t \geq 1$. Taking the negative normalized logarithm of (46), including γ_n and optimizing we find

$$\max_{s \in (0,1]} \left\{ -\frac{1}{s}\frac{1}{n} \log \gamma_n - \frac{1}{n} \log \mathbb{E}[P_{\mathrm{e}}(\mathsf{C}_n)^s]^{\frac{1}{s}} \right\}$$

$$\geq \max_{t \geq 1} \max_{\rho \in [0,1]} \left\{ -\frac{t}{n} \log \gamma_n - \rho R_n - (t-1) R_n \right.$$
$$\left. -\frac{1}{n} \log \left(\sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}')^{\frac{1}{1+\rho}} \right)^{1+\rho} \right) \right\}$$
(47)

$$\geq \max_{\rho \in [0,1]} \left\{ -\frac{1}{n} \log \gamma_n - \rho R_n \right.$$
$$\left. -\frac{1}{n} \log \left(\sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}')^{\frac{1}{1+\rho}} \right)^{1+\rho} \right) \right\},$$
(48)

where (47) is because $t = 1$ leads to the maximum.

Combining (37) and (47) and using Lemma 1 we obtain the theorem statement. ∎

Theorem 1, since based on Lemma 1, only requires γ_n to take positive values. By adding some additional constraint on γ_n, the next theorem gives a bound on the exponents of typical codes. Like Theorem 1, the following holds for channels with arbitrary alphabets or memory and pairwise-independent code ensembles.

Theorem 2: Let γ_n be a positive real-valued sequence in n such that

$$\liminf_{n \to \infty} \gamma_n = \infty,$$
(49)

$$\sum_{n=1}^{\infty} \frac{1}{\gamma_n} < \infty.$$
(50)

Then, for any channel W^n and pairwise-independent ensemble with codeword distribution Q^n it holds that:

$$\mathbb{P}\left[\liminf_{n \to \infty} E_n(\mathsf{C}_n) > \liminf_{n \to \infty} E_{\mathrm{trc,lb}}(R_n, Q^n) \right] = 1. \quad (51)$$

Proof: The proof is based on the Borel-Cantelli lemma. Such lemma is often used to upgrade convergence in probability to almost sure (a.s.) convergence. However, our theorem does not consider either kind of convergence, although the

result of the theorem is similar in spirit to going from convergence in probability to a.s. convergence.

From the assumptions and from the statement of Theorem 1 we have that $\lim_{n\to\infty} \mathbb{P}[\mathcal{A}_n(\mathsf{C}_n)] = 0$ where

$$\mathcal{A}_n(\mathcal{C}_n) = \{E_n(\mathcal{C}_n) \leq E_{\mathrm{trc,lb}}(R_n, Q^n)\}. \qquad (52)$$

Consider also the following asymptotic event [25, Sec. 1.3] [26, Sec. 2.3]:

$$\limsup_{n\to\infty} \mathcal{A}_n(\mathcal{C}_n) = \bigcap_{n\geq 1} \bigcup_{k\geq n} \mathcal{A}_k(\mathcal{C}_k) = \{\mathcal{A}_n(\mathcal{C}_n) \text{ i.o.}\}, \quad (53)$$

where "i.o." stands for "infinitely often" and denotes that events of the kind $\mathcal{A}_n(\mathcal{C}_n)$ happen for an infinite number of values of n. The event in (52) corresponds to finding a code for which the event within square brackets in (19) does not hold (i.e. the complementary event). The event in (53) corresponds to the event "infinitely many $\mathcal{A}_n(\mathcal{C}_n)$ occur". Note that in (53) the code sequence $\mathcal{C}_n, n = 1, \ldots, \infty$, is fixed.

To get an intuition on how this corresponds to the left-hand side term in (53), assume there exists a positive integer n' such that $\bigcup_{k\geq n'} \mathcal{A}_k(\mathcal{C}_k) = \emptyset$, i.e., no $\mathcal{A}_k(\mathcal{C}_k)$ occurs for $k \geq n'$. In this case less than n' events of the kind $\mathcal{A}_n(\mathcal{C}_n)$ can occur, and so only finitely many such events happen. Note also the following: consider the event $\bigcap_{n\geq 1} \mathcal{A}_n(\mathcal{C}_n)$. While the occurrence of such event implies that infinitely many $\mathcal{A}_n(\mathcal{C}_n)$ occur, it also implies that each $\mathcal{A}_n(\mathcal{C}_n), n = 1, 2, \ldots$ occur, which is not needed here. Hence the union in (53). From Theorem 1 and from (49), we have that $\mathbb{P}[\mathcal{A}_n(\mathsf{C}_n)] \leq \frac{1}{\gamma_n}$, and as a consequence that

$$\sum_{n=1}^{\infty} \mathbb{P}[\mathcal{A}_n(\mathsf{C}_n)] < \infty. \qquad (54)$$

By the Borel-Cantelli lemma [26, Sec. 2.3] and using (54) we have:

$$\mathbb{P}[\mathcal{A}_n(\mathsf{C}_n) \text{ i.o.}] = 0. \qquad (55)$$

Equation (55) is equivalent to the statement of Theorem 2. This can be formally proven noticing that equation (55) implies that the event complementary to the one within brackets has probability one and using the following identity

$$\overline{\mathcal{A}_n(\mathsf{C}_n) \text{ i.o.}} = \overline{\limsup_{n\to\infty} \mathcal{A}_n(\mathsf{C}_n)} \qquad (56)$$

$$= \liminf_{n\to\infty} \overline{\mathcal{A}_n(\mathsf{C}_n)} \qquad (57)$$

where to obtain (57) we applied twice the De Morgan's law [25, Sec. 1.2]. ∎

Some considerations are in order.

Remark 1: We use $\liminf_{n\to\infty}$ rather than $\lim_{n\to\infty}$ in (51) because on the one hand $\liminf_{n\to\infty} E_n(\mathsf{C}_n)$ necessarily exists in $\mathbb{R} \cup \{-\infty, +\infty\}$. On the other hand, for a given code the limit of the exponent might not exist even if the limit of $E_{\mathrm{ex}}^n(2R, Q^n)$ does.

Note that we used the fact that $R = \liminf_{n\to\infty} \frac{1}{n} \log M_n = \limsup_{n\to\infty} \frac{1}{n} \log M_n$, which follows from the definition of R.

Remark 2: Theorem 2 holds whether the positive sequence δ_n has a limit or not and, if it does, whether it converges or not. Let us consider the case in which it converges, that is

$$\delta_n = \frac{\hat{\rho}_n}{n} \log \gamma_n \to c, \qquad (58)$$

where $c < \infty$. The smaller c is, the tighter the bound. Hence, we are particularly interested in the case $c = 0$. Note that (58) implies $\iota_n \to 0$, since $\iota_n \leq \delta_n$. Define $\hat{\rho} = \lim_{n\to\infty} \hat{\rho}_n$. Assume that $\hat{\rho} < \infty$. Then we need that $\lim_{n\to\infty} \delta_n = \lim_{n\to\infty} \frac{1}{n} \log \gamma_n = 0$, which imposes a further constraint on γ_n. Instead, for $\hat{\rho}_n \to \infty$ we have that $\lim_{n\to\infty} \delta_n = 0$ if and only if $\hat{\rho}_n$ grows slower than $\frac{n}{\log \gamma_n}$, e.g., as $\hat{\rho}_n = \frac{\sqrt{n}}{\log \gamma_n}$. Under these assumptions, δ_n vanishes with n. Thus, γ_n must grow fast enough for the series of its reciprocal to converge, but slow enough so that the above holds. This is the case, for example, for i.i.d. codes and constant-composition codes over DMC as well as any code over finite-state channels, as we show in the next section. We analyze the case in which $\hat{\rho}_n \to \infty$ in Section III-B.

In some case it can happen that, although ι_n and δ_n go to 0 asymptotically, the statement of Lemma 1 has limited practical relevance, despite the conditions of Theorem 2 being satisfied. This is the case, for instance, in channels whose capacity is zero. As an example, we consider the quasi-static binary symmetric channel (BSC) studied in [27], where the crossover probability of the channel p is fixed for the whole transmission of a codeword, and changes randomly according to some distribution, from codeword to codeword. In this case, the ensemble-average error probability of the quasi-static BSC $\mathbb{E}[P_e(\mathsf{C}_n)]$ does not vanish with n but rather converges to a constant, the outage probability, given by $P_{\mathrm{out}}(R) = \mathbb{P}[I(p) < R_n]$, where $I(p) = 1 - h(p)$ is the mutual information of a BSC with crossover probability p, where the probability is taken with respect to the random crossover probability p. In this case, the complementary of the tail probability in (51), written in terms of error probability instead of error exponent, with $s = 1$ for simplicity, satisfies

$$\mathbb{P}\left[\liminf_{n\to\infty} P_e(\mathsf{C}_n) \geq \liminf_{n\to\infty} \gamma_n \mathbb{E}[P_e(\mathsf{C}_n)]\right]$$

$$= \mathbb{P}\left[\liminf_{n\to\infty} P_e(\mathsf{C}_n) \geq \liminf_{n\to\infty} \gamma_n \cdot P_{\mathrm{out}}(R_n)\right] \qquad (59)$$

$$= \mathbb{P}\left[\liminf_{n\to\infty} P_e(\mathsf{C}_n) \geq \infty\right] \qquad (60)$$

$$= 0, \qquad (61)$$

where we assumed that $P_{\mathrm{out}}(R_n) > 0$, in other words, that the distribution of the crossover probability p is such that the expectation in the left-hand side of (59) converges to a positive number. For such channel, Theorem 2 applies but gives a result which is not practically relevant.

There might also be cases that δ_n diverges for $R = 0$. Whether for a given pairwise code ensemble and channel this is the case or not, depends on the variance of the quantity $\log Z_n(\boldsymbol{x}, \boldsymbol{x}')$, as will be shown in Section III-B.

Remark 3: Let $E_{\mathrm{ex}}(R, Q)$ and $E_r(R, Q)$ be the asymptotic single-letter expurgated and random-coding exponents, respectively in (10) and (15), and denote by Q the limiting

distribution of Q^n. If the following limits exist

$$\lim_{n\to\infty} E_{\mathrm{ex}}^n(2R_n, Q^n) = E_{\mathrm{ex}}(2R, Q), \tag{62}$$

$$\lim_{n\to\infty} E_{\mathrm{r}}^n(R_n, Q^n) = E_{\mathrm{r}}(R, Q), \tag{63}$$

after exploiting the superadditivity of $\liminf$ in Theorem 2, it follows that

$$\mathbb{P}\Big[\liminf_{n\to\infty} E_n(\mathsf{C}_n) > \max\Big\{E_{\mathrm{ex}}(2R, Q) + R +$$
$$- \liminf_{n\to\infty} \delta_n, E_{\mathrm{r}}(R, Q) - \liminf_{n\to\infty} \iota_n\Big\}\Big] = 1. \tag{64}$$

Equation (64) holds independently on the behaviour of δ_n and ι_n and specifically on whether they (or only one of them) approaches 0 as n tends to infinity. Note that if they both diverge, (64) leads to a trivial result while if only δ_n tends to infinity while $\iota_n \to 0$ the maximum always coincides with $E_{\mathrm{r}}(R, Q)$.

Restricting to channels and ensembles for which $\delta_n \to 0$ (and thus $\iota_n \to 0$, see Remark 2), Equation (64) becomes

$$\mathbb{P}\Big[\liminf_{n\to\infty} E_n(\mathsf{C}_n) > \max\{E_{\mathrm{ex}}(2R, Q) + R, E_{\mathrm{r}}(R, Q)\}\Big] = 1. \tag{65}$$

Now, keeping the assumption that the limits exist and that (65) holds, consider the maximum R (if it exists) such that the optimal ρ leading to $E_{\mathrm{r}}(R, Q)$ is 1 and let us refer to such rate as R^*. We have: $E_{\mathrm{ex}}(R^*, Q) = E_{\mathrm{r}}(R^*, Q)$, which follows from similar arguments as in [3, Sec. 5.7]. Furthermore, since the derivatives of $E_{\mathrm{ex}}(R, Q)$ and $E_{\mathrm{r}}(R, Q)$ with respect to R are exactly $-\lambda$ and $-\rho$, respectively, we obtain that

$$\max\{E_{\mathrm{ex}}(2R, Q) + R, E_{\mathrm{r}}(R, Q)\}$$
$$= \begin{cases} E_{\mathrm{ex}}(2R, Q) + R & R \leq R^* \\ E_{\mathrm{r}}(R, Q) & R > R^*, \end{cases} \tag{66}$$

recovering the TRC exponent over all rates. We note that $E_{\mathrm{r}}(R, Q)$ might be zero above a certain rate. Optimizing over the input distribution Q, such rate indeed coincides with the channel capacity. Note also that, in case $\delta_n \to \infty$ and $\iota_n \to 0$, then $R^* = 0$, i.e., $E_{\mathrm{trc,lb}}(R_n, Q^n)$ coincides with $E_{\mathrm{r}}(R, Q)$.

Observe that Theorem 1 does not apply to code ensembles whose codeword generation is not pairwise independent, such as the random Gilbert-Varshamov (RGV) ensemble [7]. For this ensemble, it is known that the exponent of the ensemble average error probability is the maximum of the expurgated and random-coding exponents [7]. For ensembles whose average error probability achieves the expurgated exponent, like the RGV, by setting $s = 1$ in Lemma 1, we find that $\mathbb{P}\Big[E_n(\mathsf{C}_n) > E_{\mathrm{ex}}^n(R_n, Q^n) - \delta_n\Big] > 1 - \epsilon_n$, implying that the error exponent of typical codes is lower bounded by $E_{\mathrm{ex}}^n(R_n, Q^n)$. This can also be seen as a consequence of Jensen's inequality, since $\mathbb{E}[\log P_{\mathrm{e}}(\mathsf{C}_n)] \leq \log \mathbb{E}[P_{\mathrm{e}}(\mathsf{C}_n)] = E_{\mathrm{ex}}^n(R_n, Q^n)$, and thus $E_{\mathrm{trc}}(R, Q) = \lim_{n\to\infty} -\frac{1}{n}\mathbb{E}[\log P_{\mathrm{e}}(\mathsf{C}_n)] \geq E_{\mathrm{ex}}^n(R_n, Q^n)$.

A. Mismatched Decoder

While Theorem 1 and Theorem 2 assume a ML decoder, they can be easily generalized to mismatched decoding, that recovers ML as a special case. This is done by replacing $E_{\mathrm{trc,x}}(R_n, Q^n)$ and $E_{\mathrm{trc,r}}(R_n, Q^n)$ with their mismatched counterpart as follows. Let $q^n(\boldsymbol{x}, \boldsymbol{y})$ be the decoding metric, $\mathcal{D}_m$ the decision region for codeword $\boldsymbol{x}_m$ according to such metric and $\hat{m}$ the index of the codeword selected by the decoder. The mismatched decoder uses the decoding rule

$$\hat{m} = \arg\max_{m \in \{1,\ldots,M_n\}} q^n(\boldsymbol{x}_m, \boldsymbol{y}). \tag{67}$$

The choice $q^n(\boldsymbol{x}, \boldsymbol{y}) = W^n(\boldsymbol{y}|\boldsymbol{x}_m)$ recovers the ML decoder. If codeword $\boldsymbol{x}_m$ is transmitted, for all $\boldsymbol{y} \notin \mathcal{D}_m$, where $\mathcal{D}_m$ is the decoding region for message m, we have that

$$\left(\sum_{m'\neq m} \left(\frac{q^n(\boldsymbol{x}_{m'}, \boldsymbol{y})}{q^n(\boldsymbol{x}_m, \boldsymbol{y})}\right)^\tau\right)^\rho \geq 1, \tag{68}$$

for $\tau \geq 0, \rho \geq 0$. Using (68), we can bound the probability of error $P_{\mathrm{e}}(\mathcal{C}_n, m)$ and find an upper bound on the probability of error for a given code in the ensemble. Similarly to (21), we have

$$P_{\mathrm{e}}(\mathcal{C}_n) \leq \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x}_m) \left(\sum_{m'\neq m} \left(\frac{q^n(\boldsymbol{x}_{m'}, \boldsymbol{y})}{q^n(\boldsymbol{x}_m, \boldsymbol{y})}\right)^\tau\right)^\rho, \tag{69}$$

where (69) follows from (68) and from extending the sum to all $\boldsymbol{y}$.

Using (69) instead of (21) in the proof of Theorem 1 we obtain that $E_{\mathrm{trc,r}}(R_n, Q^n)$ and $E_{\mathrm{trc,x}}(R_n, Q^n)$ are now respectively given by

$$E_{\mathrm{trc,r}}(R_n, Q^n)$$
$$= \max_{\rho\in[0,1],\tau\geq 0}\Bigg\{-\frac{1}{n}\log\gamma_n - \rho R_n - \frac{1}{n}\log\Bigg(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \cdot$$
$$\cdot \sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\Bigg(\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}')\Bigg(\frac{q^n(\boldsymbol{x}', \boldsymbol{y})}{q^n(\boldsymbol{x}, \boldsymbol{y})}\Bigg)^\tau\Bigg)^\rho\Bigg)\Bigg\} \tag{70}$$

and

$$E_{\mathrm{trc,x}}(R_n, Q^n)$$
$$= \max_{\lambda\geq 1,\tau\geq 0}\Bigg\{-\frac{\lambda}{n}\log\gamma_n + R_n - \lambda 2R_n +$$
$$- \frac{\lambda}{n}\log\Bigg(\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}') \cdot$$
$$\cdot \Bigg(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\Bigg(\frac{q^n(\boldsymbol{x}', \boldsymbol{y})}{q^n(\boldsymbol{x}, \boldsymbol{y})}\Bigg)^\tau\Bigg)^{\frac{1}{\lambda}}\Bigg)\Bigg\}. \tag{71}$$

Note that, except for the term $-\frac{1}{n}\log\gamma_n$, the result in (70) recovers the expression of Gallager's mismatched decoding exponent for i.i.d. inputs [28, Eq. (7.4)]. Similarly,

subtracting R_n and $-\frac{\lambda}{n}\log\gamma_n$ from (71) and substituting $2R_n$ with R_n in the remaining expression, one recovers the expurgated exponent for the mismatched decoding i.i.d. case given in [28, Eq. (7.35)]. Using (70) and (71), Theorem 1 and Theorem 2 can be directly extended to the mismatched case.

B. Low Rates, $\hat{\rho}_n \to \infty$

For some channels and ensembles it can happen that, in the rate regime where $E_{\mathrm{trc,lb}}(R_n, Q^n) = E_{\mathrm{trc,x}}(R_n, Q^n)$, the optimal λ grows unbounded with n, i.e., $\hat{\lambda}_n \to \infty$. This is easy to see for $R = 0$, but can also happen for positive rates. We start by calculating the derivative of $E_{\mathrm{x}}^n(\lambda, Q^n)$ with respect to λ and show that it is strictly positive, except some special cases, for $n \geq 2$ and $1 \leq \lambda < \infty$. We have:

$$\frac{\partial E_{\mathrm{x}}^n(\lambda, Q^n)}{\partial \lambda}$$
$$= -\frac{1}{n}\log\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')Z^n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}} +$$
$$+ \frac{1}{n}\frac{\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x})\log\left[Z^n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}}\right]Z^n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}}}{\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')Z^n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}}}$$
$$\tag{72}$$

$$= -\frac{1}{n}\log\mathbb{E}_{Q^n}\left[Z^n(\boldsymbol{X}, \boldsymbol{X}')^{\frac{1}{\lambda}}\right] +$$
$$+ \frac{1}{n}\frac{\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')Z^n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}}\log\left[Z^n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}}\right]}{\mathbb{E}_{Q^n}\left[Z^n(\boldsymbol{X}, \boldsymbol{X}')^{\frac{1}{\lambda}}\right]}$$
$$\tag{73}$$

$$= -\frac{1}{n}\log\mathbb{E}_{Q^n}\left[Z^n(\boldsymbol{X}, \boldsymbol{X}')^{\frac{1}{\lambda}}\right] +$$
$$+ \frac{1}{n}\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} P^n(\boldsymbol{x}, \boldsymbol{x}')\log\frac{P^n(\boldsymbol{x}, \boldsymbol{x}')}{Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')} +$$
$$+ \frac{1}{n}\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} P^n(\boldsymbol{x}, \boldsymbol{x}')\log\left\{\mathbb{E}_{Q^n}\left[Z^n(\boldsymbol{X}, \boldsymbol{X}')^{\frac{1}{\lambda}}\right]\right\} \tag{74}$$

$$= \frac{1}{n}\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} P^n(\boldsymbol{x}, \boldsymbol{x}')\log\frac{P^n(\boldsymbol{x}, \boldsymbol{x}')}{Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')} \tag{75}$$

$$= \frac{1}{n}D\left(P^n\|Q^nQ^n\right) \tag{76}$$

where in (73) we substituted

$$\mathbb{E}_{Q^n}\left[Z(\boldsymbol{X}, \boldsymbol{X}')^{\frac{1}{\lambda}}\right] = \sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')Z^n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}},$$
$$\tag{77}$$

in (74) we defined

$$P^n(\boldsymbol{x}, \boldsymbol{x}') = \frac{Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')Z^n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}}}{\mathbb{E}_{Q^n}\left[Z(\boldsymbol{X}, \boldsymbol{X}')^{\frac{1}{\lambda}}\right]}, \tag{78}$$

while in (75) we used the fact that $P^n(\boldsymbol{x}, \boldsymbol{x}')$ is a probability distribution and thus its components add up to 1. As suggested by Equation (76), the derivative of $E_{\mathrm{x}}^n(\lambda, Q^n)$ is $\frac{1}{n}$ times the relative entropy between the distributions $P^n(\boldsymbol{x}, \boldsymbol{x}')$ and $Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')$, and thus is strictly positive except in some special case. Furthermore, for most channels and ensembles,

for any finite $\lambda \geq 1$ its value is bounded. This is because the relative entropy takes value ∞ only if the ratio of the distributions diverges with n for some value. Since $P(\boldsymbol{x}, \boldsymbol{x}')$ is equal to $Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')$ times another term, the derivative is infinite only if such term goes to 0 fast enough so that the double sum grows asymptotically faster than n.

We now study the behaviour of $E_{\mathrm{x}}^n(\lambda, Q^n)$ for $\lambda \to \infty$. Since for all n the derivative is non-negative, $E_{\mathrm{x}}^n(\lambda, Q^n)$ can either converge to a constant or diverge. By calculating the limit we obtain

$$\lim_{\lambda\to\infty} E_{\mathrm{x}}^n(\lambda, Q^n) = -\frac{1}{n}\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')\log Z_n(\boldsymbol{x}, \boldsymbol{x}').$$
$$\tag{79}$$

Note that the right-hand side of (79) can be ∞ in some case depending on the distribution of $\log Z_n(\boldsymbol{X}, \boldsymbol{X}')$.

Now let us consider expression (10) and study the cases in which, for $R_n > 0$, the optimal $\hat{\lambda}_n$ grows unbounded as $n \to \infty$. This happens for all rates $R_n < R_\infty(Q^n)$, where $R_\infty(Q^n) > 0$ is a quantity that can be found following a similar reasoning as in [3, Sec. 5.7]. Let us study the R-axis intercept of the following linear function $E_{\mathrm{x}}^n(\lambda, Q^n) + R_n - \lambda 2R_n$, given by

$$R_n = \frac{E_{\mathrm{x}}^n(\lambda, Q^n)}{2\lambda - 1}. \tag{80}$$

Using (75), (77), (78), and taking the limit for $\lambda \to \infty$ we have

$$R_\infty(Q^n)$$
$$= \lim_{n\to\infty} R_n \tag{81}$$
$$= \lim_{n\to\infty} -\frac{1}{2}\frac{1}{n}\log\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')\mathbb{1}\{Z_n(\boldsymbol{x}, \boldsymbol{x}') > 0\}$$
$$\tag{82}$$
$$= \lim_{n\to\infty} -\frac{1}{2}\frac{1}{n}\log\mathbb{P}\left[Z_n(\boldsymbol{x}, \boldsymbol{x}') > 0\right]. \tag{83}$$

Expression (83) is (one half) the exponent of the probability that the Bhattacharyya bound on the pairwise error probability be positive. For all rates smaller than $R_\infty(Q^n)$ as given in (83), the quantity $E_{\mathrm{x}}^n(\lambda_n, Q^n) - \lambda_n 2R_n + R_n$ is infinite, which implies that $E_{\mathrm{x}}^n(\lambda_n, Q^n)$ is given by (79) because $\hat{\lambda}$ is ∞. Notice that this is achieved using a $\hat{\lambda}_n$ that tends to infinity as n grows. In order to assess the behaviour of the TRC exponent we need to study the δ_n in Definition 9 as $\hat{\lambda}_n \to \infty$. We recall that $\delta_n = \frac{\hat{\lambda}_n}{n}\log\gamma_n$. Repeating the derivation of $R_\infty(Q^n)$ including δ_n in the calculation, Equation (80) becomes (ignoring the -1 at the denominator which has no impact asymptotically):

$$R_n = \frac{E_{\mathrm{x}}^n(\lambda, Q^n) - \lambda\frac{\log\gamma}{n}}{2\lambda} = \frac{E_{\mathrm{x}}^n(\lambda, Q^n)}{2\lambda} - \frac{\log\gamma}{2n}. \tag{84}$$

Taking the limit in (84) we obtain again $R_\infty(Q^n)$ as in (83). If $R_\infty(Q^n) > 0$, for all rates $R_n < R_\infty(Q^n)$ the TRC exponent is infinite. Note that in such case it can happen that $\delta_n \to \infty$, as λ_n must grow asymptotically like $E_{\mathrm{x}}^n(\lambda, Q^n)$, which could be actual faster than $\frac{n}{\log\gamma_n}$. However, also in

such case the statement of Theorem 2 would give a non-trivial result, as $E_x^n(\lambda, Q^n)$ grows faster than δ_n. The $R_\infty(Q^n)$ derived here is closely related to the quantity R_x^∞ derived in [3, Sec. 5.7] for the DMC and is, in fact, equal to $1/2$ times the equivalent of R_x^∞ for general channels and pairwise-independent code ensembles. Similarly to $R_{x,\infty}$, the quantity $2R_\infty(Q^n)$ is a lower bound on the zero-error capacity for general channels. This implies that for all code ensembles and channels considered in the present section for which the zero-error capacity is 0, $R_\infty(Q^n) = 0$ as well.

Let us now consider the case in which $R_\infty(Q^n) = 0$. Expression (83) implies $R_\infty(Q^n) = 0$ whenever the probability that the Bhattacharyya bound on the codeword transition probability decreases sub-exponentially. To understand what happens to the TRC exponent in such case, we need to check both (79), which is the average exponent of the bound, and the behaviour of δ_n. Note that expression (79) does not necessarily go to ∞ when $R_\infty(Q^n) = 0$. For strictly zero rates, namely when $\lim_{n\to\infty} \frac{1}{n} \log M_n = 0$, the largest exponent in (10) is achieved for $\hat{\lambda}_n \to \infty$. We then consider the dependence on n in the rate by using the definition $R_n = \frac{1}{n} \log M_n$ in (12) and let $\hat{\lambda}_n$ be

$$\hat{\lambda}_n = \arg\max_{\lambda \geq 1} \left\{ E_x^n(\lambda, Q^n) - \frac{2\lambda}{n} \log M_n \right\}. \quad (85)$$

In order to characterize how $\hat{\lambda}_n$ grows as $n \to \infty$, we find the Taylor series expansion of $E_x^n(\lambda, Q^n)$ around $\lambda \to \infty$, yielding

$$E_x^n(\lambda, Q^n) = \nu_0^n(Q^n) - \frac{\nu_1^n(Q^n)}{\lambda} + O(\lambda^{-2}), \quad (86)$$

where

$$\nu_0^n(Q^n) = -\frac{1}{n} \mathbb{E}[\log Z_n(\boldsymbol{X}, \boldsymbol{X}')] \quad (87)$$

and

$$\nu_1^n(Q^n) = \frac{1}{2n} \text{Var}[\log Z_n(\boldsymbol{X}, \boldsymbol{X}')]. \quad (88)$$

Then, it follows that $\hat{\lambda}_n$ grows as

$$\hat{\lambda}_n = \sqrt{\frac{\nu_1^n(Q^n)}{2\frac{\log M_n}{n} + \frac{\log \gamma_n}{n}}} + \sigma_n, \quad (89)$$

where the two terms at the denominator vanish with n, and also $\sigma_n \to 0$. From (89) it follows that δ_n defined in (18) satisfies $\lim_{n\to\infty} \delta_n = 0$ for $R = 0$ if $\frac{1}{n} \text{Var}[Z_n(\boldsymbol{X}, \boldsymbol{X}')]$ grows slower than $\sqrt{\frac{n}{\log \gamma_n}}$. Note that $\text{Var}[\log Z_n(\boldsymbol{X}, \boldsymbol{X}')]$ depends both on the channel and the specific code ensemble. The parameter $\nu_0^n(Q^n)$ in (87) is the same as the expression in (79). Note that the analysis on the growth rate of $\hat{\lambda}_n$ just carried out applies to all cases in which $\hat{\lambda}_n \to \infty$. Thus, apart from the case $R = 0$, it also holds for $0 < R < R_\infty(Q^n)$. In such case the denominator of (89) does not vanish with n. This implies that $\text{Var}[\log Z_n(\boldsymbol{X}, \boldsymbol{X}')] \to \infty$ faster than n. The term ν_1^n in (88) characterizes the backoff of the error exponent (86) from its limiting value as $\lambda \to \infty$.

IV. MEMORYLESS CHANNELS

In this section, we specialize Theorem 2 to memoryless channels with i.i.d., constant-composition and cost-constrained ensembles. As a by-product, we provide a direct dual-domain derivation of the expurgated exponent for constant-composition-codes [6], as an alternative to that in [29].

A. i.i.d. Ensemble

Let us consider the i.i.d. ensemble with distribution

$$Q_{\text{iid}}^n(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i). \quad (90)$$

Using (90) in (10), we recover the single-letter version of the Gallager's E_x-function [3, Eq. (5.7.12)], namely

$$E_x^{\text{iid}}(\lambda, Q) = -\lambda \log \sum_x \sum_{x'} Q(x) Q(x') Z(x,x')^{\frac{1}{\lambda}}, \quad (91)$$

where $Z(x, x')$ is the single-letter Bhattacharyya coefficient, and the i.i.d. expurgated exponent in (11) becomes

$$E_{\text{ex}}^{\text{iid}}(R, Q) = E_x^{\text{iid}}(\hat{\lambda}, Q) - \hat{\lambda}R. \quad (92)$$

When $R > 0$, it follows that the optimal parameter (12) does not depend on n, its limit exists and reads

$$\hat{\lambda} = \arg\max_{\lambda \geq 1} \left\{ E_x^{\text{iid}}(\lambda, Q) - \lambda 2R \right\}. \quad (93)$$

In all cases for which $\hat{\lambda} < \infty$, the condition $\lim_{n\to\infty} \delta_n = 0$ of the theorem is satisfied. For all cases in which $\hat{\rho}$ is infinite we can specialize the discussion in Section III-B to the i.i.d. case. For strictly zero rates, namely when $\lim_{n\to\infty} \frac{1}{n} \log M_n = 0$, the largest exponent in (92) is achieved for $\hat{\lambda} \to \infty$. Then, from (89), it follows that $\hat{\lambda}_n$ grows as

$$\hat{\lambda}_n = \sqrt{\frac{n \cdot \nu_1(Q)}{2 \log M_n + \log \gamma_n}} + \sigma_n, \quad (94)$$

where

$$\nu_1(Q) = \frac{1}{2} \text{Var} \left[\log Z(x, x') \right] \quad (95)$$

is constant in n while σ_n is a term that vanishes as $n \to \infty$, implying that δ_n defined in (18) satisfies $\lim_{n\to\infty} \delta_n = 0$ as in the $R > 0$ case.

The case in which $\hat{\lambda}_n$ grows unbounded for positive rates is obtained by replacing $R_\infty(Q^n)$ (after optimization over Q^n) with $\frac{1}{2} R_x^\infty$ [3] in Section III-B. As mentioned for the general case, also for the DMC the δ_n can, in principle, diverge when $R < R_\infty(Q^n)$, and yet $E_{\text{trc,lb}}(R_n, Q^n) \to \infty$.

According to Theorem 1, with probability approaching one as $n \to \infty$, an i.i.d. code $\mathcal{C}_n^{\text{iid}}$ randomly generated from the ensemble (90) has an error exponent (3) satisfying

$$E_n(\mathcal{C}_n^{\text{iid}}) \geq E_{\text{ex}}^{\text{iid}}(2R_n, Q) + R_n - \delta_n \quad (96)$$

with vanishing δ_n. Particularizing (92) to the BSC case, it follows that the bound (96) coincides with the TRC exponent given by Barg and Forney in [8].

Remark 4: Let us denote the smallest rate for which $\hat{\lambda} = 1$ with R^* and recall that the expurgated exponent and the random-coding exponent coincide for such rate [3, pg. 154]. Similarly as in Remark 3, it can be shown that, for all cases in which $\delta_n \to 0$, Theorem 2 implies that the achievable typical random-coding exponent is lower bounded as follows:

$$E_{\mathrm{trc}}(R, Q) \geq \begin{cases} E_{\mathrm{ex}}(2R, Q) + R & 0 \leq R \leq \frac{R*}{2} \\ E_{\mathrm{r}}(R, Q) & R > \frac{R*}{2}. \end{cases} \quad (97)$$

Note that $E_{\mathrm{r}}(R, Q)$ might become zero above a certain rate. If a maximization with respect to the input distribution is carried out, such rate coincides with the capacity of the channel. This lower bound is known to be tight for constant-composition ensembles [10] and coincides with the one derived in [8] for the BSC.

B. Constant-Composition Ensemble

For every n, let $\hat{Q}_n$ be a type, or empirical distribution, such that $\|\hat{Q}_n - Q\|_\infty \leq \frac{1}{n}$ where $\|P\|_\infty = \max_x P(x)$. Then, the constant-composition ensemble has codeword distribution

$$Q_{\mathrm{cc}}^n(\boldsymbol{x}) = \frac{1}{|\mathcal{T}^n(\hat{Q}_n)|} \mathbb{1}\{\boldsymbol{x} \in \mathcal{T}^n(\hat{Q}_n)\}, \quad (98)$$

where $\mathcal{T}_n(\hat{Q}_n)$ is the type class, i.e., the set of all sequences of length n with empirical distribution $\hat{Q}_n$, while $\mathbb{1}\{.\}$ is the indicator function.

For the constant-composition ensemble (98), the normalized multi-letter Gallager's expurgated function $E_{\mathrm{x}}^n(\lambda, Q^n)$ reads

$$E_{\mathrm{x}}^n(\lambda, Q^n) = -\frac{1}{n} \log \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}') \prod_{i=1}^n Z(x_i, x_i')^{\frac{1}{\lambda}} \right)^\lambda. \quad (99)$$

In (99), we have used the fact that the channel is memoryless to express $Z_n(\boldsymbol{x}, \boldsymbol{x}')$ as a product, and we have also changed the order of the summations over $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Since all codewords $\boldsymbol{x}'$ have the same probability $\frac{1}{|\mathcal{T}^n(\hat{Q}_n)|}$, the summation over $\boldsymbol{x}'$ in (99) for a fixed constant-composition sequence $\boldsymbol{x}$ satisfies

$$\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}') \prod_{i=1}^n Z(x_i, x_i')^{\frac{1}{\lambda}}$$

$$= \frac{1}{|\mathcal{T}^n(\hat{Q}_n)|} \sum_{\boldsymbol{x}'} \prod_{i=1}^n Z(x_i, x_i')^{\frac{1}{\lambda}} \quad (100)$$

$$\leq \min_{\bar{P}} \left\{ (n+1)^{|\mathcal{X}|-1} e^{nD(\hat{Q}_n \| \bar{P})} \prod_{i=1}^n \sum_{x'} \bar{P}(x') Z(x_i, x')^{\frac{1}{\lambda}} \right\}, \quad (101)$$

where the upper bound in (101) follows by identifying $g_i(x_i) = Z(x_i, x_i')^{\frac{1}{\lambda}}$ in [30, Eq. (2.4)], $\bar{P}$ is an auxiliary probability distribution and $D(Q\|\bar{P})$ is the relative entropy

between distributions Q and $\bar{P}$. Using the upper bound (101) into (99) and arranging terms, we obtain

$$E_{\mathrm{x}}^n(\lambda, Q^n)$$

$$\geq \max_{\bar{P}} \left\{ -\frac{1}{n} \log \left((n+1)^{|\mathcal{X}|-1} e^{nD(\hat{Q}_n \| \bar{P})} \cdot \right. \right.$$

$$\left. \left. \cdot \frac{1}{|\mathcal{T}^n(\hat{Q}_n)|} \sum_{\boldsymbol{x}} \prod_{i=1}^n \sum_{x'} \bar{P}(x') Z(x_i, x')^{\frac{1}{\lambda}} \right)^\lambda \right\} \quad (102)$$

$$= \max_{\bar{P}} \left\{ -\frac{1}{n} \log \left((n+1)^{|\mathcal{X}|-1} e^{nD(\hat{Q}_n \| \bar{P})} \prod_x \left(\sum_{x'} \bar{P}(x') \cdot \right. \right. \right.$$

$$\left. \left. \left. \cdot Z(x, x')^{\frac{1}{\lambda}} \right)^{n\hat{Q}_n(x)} \right)^\lambda \right\}, \quad (103)$$

where (103) follows since the sum over $\boldsymbol{x}$ in (102) has exactly $|\mathcal{T}^n(\hat{Q}_n)|$ terms and from elementary properties of constant-composition sequences. Next, we rewrite the maximization over $\bar{P}$ as maximization over an auxiliary function $a(x)$ [29] satisfying $\bar{P}(x) = \hat{Q}_n(x) e^{\frac{a(x)}{\lambda}}$. Using this in (103) we find that

$$E_{\mathrm{x}}^n(\lambda, Q^n)$$

$$\geq \max_{a(\cdot)} \left\{ -\frac{1}{n} \log \left((n+1)^{|\mathcal{X}|-1} e^{-\frac{n}{\lambda} \sum_x \hat{Q}_n(x) a(x)} \cdot \right. \right.$$

$$\left. \left. \cdot \prod_x \left(\sum_{x'} \hat{Q}_n(x') e^{\frac{a(x')}{\lambda}} Z(x, x')^{\frac{1}{\lambda}} \right)^{n\hat{Q}_n(x)} \right)^\lambda \right\} \quad (104)$$

$$= -\frac{\lambda(|\mathcal{X}|+1)}{n} \log(n+1) + \max_{a(\cdot)} \left\{ -\lambda \log \prod_x \left(\sum_{x'} \hat{Q}_n(x') \cdot \right. \right.$$

$$\left. \left. \cdot \left(Z(x, x') e^{a(x')-a(x)} \right)^{\frac{1}{\lambda}} \right)^{\hat{Q}_n(x)} \right\}. \quad (105)$$

We observe that the the first term in the r.h.s. of (105) vanishes with n. Using that $\|\hat{Q}_n - Q\|_\infty \leq \frac{1}{n}$, and further simplifying the result, we obtain the constant-composition version of the Gallager's E_{x}-function, namely

$$E_{\mathrm{x}}^{\mathrm{cc}}(\lambda, Q) = \max_{a(\cdot)} \left\{ -\lambda \sum_x Q(x) \log \sum_{x'} Q(x') \Big(Z(x, x') \cdot \right.$$

$$\left. \cdot e^{a(x')-a(x)} \Big)^{\frac{1}{\lambda}} \right\}. \quad (106)$$

According to Theorem 1, with probability approaching one as $n \to \infty$, a constant-composition code $\mathcal{C}_n^{\mathrm{cc}}$ randomly generated from the ensemble (98) has an error exponent satisfying

$$E_n(\mathcal{C}_n^{\mathrm{cc}}) \geq E_{\mathrm{ex}}^{\mathrm{cc}}(2R, Q) + R, \quad (107)$$

where

$$E_{\mathrm{ex}}^{\mathrm{cc}}(R, Q) = E_{\mathrm{x}}^{\mathrm{cc}}(\hat{\lambda}, Q) - \hat{\lambda} R \quad (108)$$

is the constant-composition version of the expurgated exponent with optimal parameter

$$\hat{\lambda} = \arg\max_{\lambda \geq 1} \{ E_{\mathrm{x}}^{\mathrm{cc}}(\lambda, Q) - \lambda 2R \}, \quad (109)$$

coinciding with the expression in [10]. Similarly to what seen for the i.i.d. case, we need to address the case in which $\hat{\lambda}_n$ diverges as n grows. For the zero-rate case we find that the optimal $\hat{\lambda}_n$ grows as

$$\hat{\lambda}_n = \sqrt{\frac{n \cdot \nu_1(Q)}{2 \log M_n + \log \gamma_n}} + \sigma_n, \qquad (110)$$

where

$$\nu_1(Q) = \max_{a(\cdot)} \frac{1}{2} \mathrm{Var} \left[\log \left(Z(x, x') e^{a(x') - a(x)} \right) \right], \quad (111)$$

implying that $\delta_n \to 0$.

The case in which $\hat{\lambda}_n$ grows unbounded for positive rates is obtained in a similar way as for the i.i.d. case by using the equivalent of R_{x}^∞ for the constant-composition case, which can be derived in a similar way as done for the general case in Section III-B and the i.i.d. case in [3, Sec. 5.7].

At high rates the typical error exponent coincides with the random-coding exponent for constant-composition ensembles [31]

$$E_{\mathrm{r}}^{\mathrm{cc}}(Q, R) = \max_{\overline{Q}} \max_{0 \le \rho \le 1} E_0(\overline{Q}, \rho) - (1 + \rho) D(Q \| \overline{Q}) - \rho R, \qquad (112)$$

where $E_0(\overline{Q}, \rho)$ is Gallager's E_0 function for the i.i.d. ensemble while $D(Q \| \overline{Q})$ is the relative entropy between the distributions Q and $\overline{Q}$.

C. Cost-Constrained Ensemble

The cost-constrained ensemble is described by a set of symbol-wise cost functions and an i.i.d. input distribution $Q^n(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i)$ that induce the pairwise-independent codeword distribution:

$$Q_{\mathrm{cost}}^n(\boldsymbol{x}) = \frac{1}{\mu_n} \prod_{i=1}^n Q(x_i) \mathbb{1}\{\boldsymbol{x} \in \mathcal{D}_n\}, \qquad (113)$$

where μ_n is a normalization factor and $\mathcal{D}_n$ is the set of input sequences satisfying the cost constraints given by

$$\mathcal{D}_n = \left\{ \boldsymbol{x} : \left| \frac{1}{n} \sum_{i=1}^n a_\ell(x_i) - \phi_\ell \right| \le \frac{\delta}{n}, \ell = 1, \ldots, L \right\} \qquad (114)$$

for some $\delta > 0$ and $\phi_\ell = \sum_x Q(x) a_\ell(x)$ where $\{a_\ell(x)\}_{\ell=1}^L$ are auxiliary cost functions and can be optimized to obtain an improved exponent with respect to the i.i.d. case. A system cost of the form $\frac{1}{n} \sum_{i=1}^n c(x_i) \le \Gamma$ can be included in the analysis, except that no optimization over the function $c(x)$ itself can be done. This ensemble [29] is a generalization of the one in [3, Sec. 7.3], which considers a single constraint. As done for the i.i.d. and the constant-composition ensembles, we specialize the result of Theorem 2 to the cost constrained case by upper bounding $E_{\mathrm{trc,x}}(R_n, Q^n)$ and $E_{\mathrm{trc,r}}(R_n, Q^n)$. The key advantage of this ensemble is that it allows to obtain the same performance of constant-composition codes for channels with possibly continuous alphabets.

Let us start with $E_{\mathrm{trc,x}}(R_n, Q^n)$. Plugging (113) into (11) we obtain

$$E_{\mathrm{x}}^n(\lambda, Q_{\mathrm{cost}}^n) = -\frac{1}{n} \log \left(\frac{1}{\mu_n^2} \sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \cdot \right.$$
$$\left. \cdot \mathbb{1}\{\boldsymbol{x} \in \mathcal{D}_n\} \mathbb{1}\{\boldsymbol{x}' \in \mathcal{D}_n\} Z_n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}} \right)^\lambda. \qquad (115)$$

Under mild assumptions on the cost functions, namely and $\mathbb{E}_Q[a_\ell(X)^2] \le \infty$ for all $\ell = 1, 2, \ldots, L$, the normalization constant satisfies [29, Prop. 2], [32] $\lim_{n \to \infty} \frac{1}{n} \mu_n = 0$, implying that, for n large,

$$E_{\mathrm{x}}^n(\lambda, Q_{\mathrm{cost}}^n) = -\frac{1}{n} \log \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \cdot \right.$$
$$\left. \cdot \mathbb{1}\{\boldsymbol{x} \in \mathcal{D}_n\} \mathbb{1}\{\boldsymbol{x}' \in \mathcal{D}_n\} Z_n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}} \right)^\lambda. \qquad (116)$$

We next find an upper bound on the indicator function in (116). Define $a_\ell^n(\boldsymbol{x}) = \sum_{i=1}^n a_\ell(x_i)$. For any real-valued number r_ℓ, we have

$$\mathbb{1}\{\boldsymbol{x} \in \mathcal{D}_n\} = \prod_{\ell=1}^L \mathbb{1}\{|a_\ell^n(\boldsymbol{x}) - n\phi_\ell| \le \delta\} \qquad (117)$$

$$\le \prod_{\ell=1}^L e^{r_\ell(a_\ell^n(\boldsymbol{x}) - n\phi_\ell)} e^{|r_\ell|\delta} \qquad (118)$$

$$= e^{\sum_{\ell=1}^L r_\ell(a_\ell^n(\boldsymbol{x}) - n\phi_\ell) + |r_\ell|\delta}, \qquad (119)$$

where in (118) we upper bounded each indicator function using (114) for any real number r_ℓ [29, Eq. (56)], which is independent of n. Using (119) we bound the product of indicator functions in (115) as:

$$\mathbb{1}\{\boldsymbol{x} \in \mathcal{D}_n\} \mathbb{1}\{\boldsymbol{x}' \in \mathcal{D}_n\}$$
$$\le e^{\sum_{\ell=1}^L r_\ell(a_\ell^n(\boldsymbol{x}) - n\phi_\ell) + |r_\ell|\delta} e^{\sum_{\ell=1}^L \overline{r}_\ell(a_\ell^n(\boldsymbol{x}') - n\phi_\ell) + |\overline{r}_\ell|\delta} \qquad (120)$$

$$= \frac{e^{\sum_{\ell=1}^L r_\ell(a_\ell^n(\boldsymbol{x}) - n\phi_\ell)}}{e^{\sum_{\ell=1}^L r_\ell'(a_\ell^n(\boldsymbol{x}') - n\phi_\ell)}} e^{\delta \sum_{\ell=1}^L (|r_\ell| + |r_\ell'|)} \qquad (121)$$

where in (121) we renamed $\overline{r}$ as $-r'$. Since r_ℓ is a real-valued variable which does not depend on n, we can bound the argument of the logarithm in (116) using (121):

$$\left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \mathbb{1}\{\boldsymbol{x} \in \mathcal{D}_n\} \mathbb{1}\{\boldsymbol{x}' \in \mathcal{D}_n\} \cdot \right.$$
$$\left. \cdot Z_n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}} \right)^\lambda \qquad (122)$$

$$\le \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \frac{e^{\sum_{\ell=1}^L r_\ell'(a_\ell^n(\boldsymbol{x}') - n\phi_\ell)}}{e^{\sum_{\ell=1}^L r_\ell(a_\ell^n(\boldsymbol{x}) - n\phi_\ell)}} \cdot \right.$$
$$\left. \cdot e^{\delta \sum_{\ell=1}^L (|r_\ell'| + |r_\ell|)} Z_n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}} \right)^\lambda \qquad (123)$$

$$\doteq \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \frac{e^{\sum_{\ell=1}^L r_\ell'(a_\ell^n(\boldsymbol{x}') - n\phi_\ell)}}{e^{\sum_{\ell=1}^L r_\ell(a_\ell^n(\boldsymbol{x}) - n\phi_\ell)}} Z_n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}} \right)^\lambda, \qquad (124)$$

where in (124) we used the fact that $e^{\delta \sum_{\ell=1}^{L}(|r_\ell|+|r_\ell'|)}$ is independent of n while $Q^n(\boldsymbol{x}) = \prod_{i=1}^{n} Q(x_i)$ is an input distribution that satisfies the conditions of [29, Prop. 2] as well as the system cost constraint . From (124) we can bound $E_{\mathrm{x}}^n(\lambda, Q_{\mathrm{cost}}^n)$ as follows:

$$
\begin{aligned}
&E_{\mathrm{x}}^n(\lambda, Q_{\mathrm{cost}}^n) \\
&\geq E_{\mathrm{x}}^n(\lambda, Q^n, \{a_\ell\}, \{r_\ell\}, \{r_\ell'\}) \\
&= -\frac{1}{n} \log \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \frac{e^{\sum_{\ell=1}^{L} r_\ell'(a_\ell^n(\boldsymbol{x}')-n\phi_\ell)}}{e^{\sum_{\ell=1}^{L} r_\ell(a_\ell^n(\boldsymbol{x})-n\phi_\ell)}} \cdot \right. \\
&\qquad\qquad\qquad \left. \cdot Z_n(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{\lambda}} \right)^\lambda . \quad (125)
\end{aligned}
$$

Expanding each term as single-letter product and optimizing over the auxiliary variables, we obtain:

$$
\begin{aligned}
&E_{\mathrm{x}}^n(\lambda, Q_{\mathrm{cost}}^n) \\
&\geq E_{\mathrm{x}}(\lambda, Q) \\
&= \sup_{\{r_\ell\}_{\ell=1}^L, \{r_\ell'\}_{\ell=1}^L, \{a_\ell(.)\}_{\ell=1}^L} \left\{ -\lambda \log \left(\sum_x \sum_{x'} Q(x)Q(x') \cdot \right.\right. \\
&\qquad\qquad \left.\left. \cdot \frac{e^{\sum_{\ell=1}^{L} r_\ell'(a_\ell(x')-\phi_\ell)}}{e^{\sum_{\ell=1}^{L} r_\ell(a_\ell(x)-\phi_\ell)}} Z(x,x')^{\frac{1}{\lambda}} \right) \right\} .
\end{aligned}
$$
$$(126)$$

From (126), the bound $E_{\mathrm{trc,x}}(R_n, Q^n)$ follows. As a side remark we point out that the expurgated of the i.i.d. and the constant-composition ensembles can be recovered from (126). The former is recovered by setting $r_\ell = r_\ell' = 0$ in (126), while the latter is obtained by setting $L = 1$, $r_1' = 1$ and $r_1 = 0$. This differs from the similar result in [29], where the constant-composition exponent is obtained with $L = 2$, due to the fact that [29] considers a mismatched decoder rather than a ML one as in our case.

Let us now consider the case in which $\hat{\lambda}_n \to \infty$ as n grows. For the zero-rate case we find from (89) that

$$
\hat{\lambda}_n = \sqrt{\frac{n \cdot \nu_1(Q)}{2 \log M_n + \log \gamma_n}} + \sigma_n, \quad (127)
$$

where now the term $\nu_1(Q)$ is given by

$$
\nu_1(Q) = \sup_{\{r_\ell\}_{\ell=1}^L, \{r_\ell'\}_{\ell=1}^L, \{a_\ell(.)\}_{\ell=1}^L} \left\{ \frac{1}{2} \mathrm{Var} \left[\right.\right.
$$
$$
\left.\left. \log \left(\frac{e^{\sum_{\ell=1}^{L} r_\ell'(a_\ell(x')-\phi_\ell)}}{e^{\sum_{\ell=1}^{L} r_\ell(a_\ell(x)-\phi_\ell)}} Z(x,x') \right) \right] \right\}, \quad (128)
$$

implying that $\delta_n \to 0$.

The case in which $\hat{\lambda}_n$ grows unbounded for positive rates is obtained in a similar way as for the i.i.d. and the constant-composition ensembles by using the equivalent of R_{x}^∞ for the cost constraint case.

As for the bound $E_{\mathrm{trc,r}}(R_n, Q^n)$, it can be obtained by plugging (113) into (15), using the bound on the indicator function given in (119) and following a similar reasoning as for $E_{\mathrm{x}}^n(\lambda, Q_{\mathrm{cost}}^n)$ to obtain [29], [31]

$$
E_0^n(\rho, Q_{\mathrm{cost}}^n)
$$

$$
\geq E_0^n(\rho, Q)
$$
$$
\begin{aligned}
= \sup_{\{r_\ell\}_{\ell=1}^L, \{r_\ell'\}_{\ell=1}^L, \{a_\ell(.)\}_{\ell=1}^L} & \log \left(\sum_y \left(\sum_x Q(x) \cdot \right.\right. \\
& \left.\left. \cdot e^{\sum_{\ell=1}^{L} r_\ell(a_\ell(x)-\phi_\ell)} W(y|x)^{\frac{1}{1+\rho}} \right)^{1+\rho} \right), \quad (129)
\end{aligned}
$$

from which the bound on the TRC exponent at high rates $E_{\mathrm{trc,r}}(R_n, Q^n)$ follows.

V. FINITE-STATE CHANNELS

In this section, we particularize the result in Theorem 2 to the case of finite-state channels. In particular, we derive the two lower bounds on the exponents $E_{\mathrm{trc,x}}(R_n, Q^n)$ and $E_{\mathrm{trc,r}}(R_n, Q^n)$ in (51).

We consider finite-state channels with A states $\{1, 2, \ldots, A\}$. The statistical behaviour of the channel is described by a joint probability measure linking the current channel state and output to the current input and previous state $W(y_t, s_t | x_t, s_{t-1})$. Similarly to [3, Sec. 4.6], our results are based on the probability of the channel output $\boldsymbol{y}$, given the channel input $\boldsymbol{x}$ and initial state s_0, which can be obtained by summing over all the possible state sequences $(s_1, \ldots, s_n)$ using the following recursive expression,

$$
W^n(\boldsymbol{y}|\boldsymbol{x}, s_0) = \sum_{s_n=1}^{A} W_n(\boldsymbol{y}, s_n | \boldsymbol{x}, s_0), \quad (130)
$$

where the probability $W_n(\boldsymbol{y}, s_n | \boldsymbol{x}, s_0)$ can be calculated using the recursion

$$
\begin{aligned}
&W_t(\boldsymbol{y}_t, s_t | \boldsymbol{x}_t, s_0) = \\
&\sum_{s_{t-1}=1}^{A} W(y_t, s_t | x_t, s_{t-1}) W_{t-1}(\boldsymbol{y}_{t-1}, s_{t-1} | \boldsymbol{x}_{t-1}, s_0), \quad (131)
\end{aligned}
$$

for $t = 1, \ldots, n$. In (131), $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are respectively the channel input and output sub-sequences given by $\boldsymbol{x}_t = (x_1, \ldots, x_t)$ and $\boldsymbol{y}_t = (y_1, \ldots, y_t)$, while for $t = 1$ we have the initial relation $W_1(y_1, s_1 | x_1, s_0) = W(y_1, s_1 | x_1, s_0)$. Equation (130) represents a family of channel transition probabilities, indexed by the initial state s_0, that also encompasses compound channels since s_0 can be considered as an index that determines the channel from the compound set. We study the error probability of ensemble of codes using a mismatched decoder that employs a decoding metric that averages over the initial state assuming an equiprobable state distribution, as in [3],

$$
W^n(\boldsymbol{y}|\boldsymbol{x}) = \sum_{s_0=1}^{A} \frac{1}{A} W^n(\boldsymbol{y}|\boldsymbol{x}, s_0). \quad (132)
$$

A. Lower Bound on $E_{\mathrm{trc,x}}(R_n, Q^n)$ for Finite-State Channels

From Lemma 1 and the error probability bound in (35), we have that with high probability, a code drawn randomly from a pairwise-independent ensemble used over a FSC and

averaged over all possible states has an error probability bounded by

$$P_e(\mathcal{C}_n)$$
$$\leq \frac{1}{M_n} \left(\gamma_n M_n(M_n-1)\right)^\lambda \left(\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')\cdot \right.$$
$$\left. \cdot\left(\sum_{\boldsymbol{y}} \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x}')}\right)^{\frac{1}{\lambda}}\right)^\lambda\right) \tag{133}$$

$$\leq \frac{1}{M_n} \left(\gamma_n M_n(M_n-1)\right)^\lambda \left(\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')\cdot \right.$$
$$\left. \cdot\left(\sum_{\boldsymbol{y}} \sqrt{\sum_{s_0=1}^A \frac{1}{A}W^n(\boldsymbol{y}|\boldsymbol{x},s_0)\sum_{s_0'=1}^A \frac{1}{A}W^n(\boldsymbol{y}|\boldsymbol{x}',s_0')}\right)^{\frac{1}{\lambda}}\right)^\lambda\right), \tag{134}$$

where (134) follows from (132). From (134) it is possible to find a bound that holds for any given initial state, useful in case a distribution over such state is unknown. Referring to the error probability for code $\mathcal{C}_n$ given an initial state $\bar{s}_0$ as $P_e(\mathcal{C}_n,\bar{s}_0)$, we have the bounds

$$P_e(\mathcal{C}_n) = \sum_{\bar{s}_0} q(\bar{s}_0)P_e(\mathcal{C}_n,\bar{s}_0) \tag{135}$$
$$\geq \max_{\bar{s}_0} q(\bar{s}_0)P_e(\mathcal{C}_n,\bar{s}_0) \tag{136}$$
$$\geq \frac{1}{A}P_e(\mathcal{C}_n,\bar{s}_0). \tag{137}$$

From (137) we obtain

$$P_e(\mathcal{C}_n,\bar{s}_0) \leq AP_e(\mathcal{C}_n), \tag{138}$$

a bound that holds independently of the initial state distribution or on whether such distribution exists or not. Plugging (134) into (138) we get an upper bound on the average error probability for a given initial state which does not depend on the initial state distribution. From Theorem 1, keeping the dependency on the initial state into account, we have that, with high probability, for any initial state s_0:

$$P_e(\mathcal{C}_n,s_0)$$
$$\leq \frac{A}{M_n} \left(\gamma_n M_n(M_n-1)\right)^\lambda \left(\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')\cdot \right.$$
$$\left. \cdot\left(\sum_{\boldsymbol{y}} \sqrt{\sum_{s_0=1}^A \frac{1}{A}W^n(\boldsymbol{y}|\boldsymbol{x},s_0)\sum_{s_0'=1}^A \frac{1}{A}W^n(\boldsymbol{y}|\boldsymbol{x}',s_0')}\right)^{\frac{1}{\lambda}}\right)^\lambda\right) \tag{139}$$

$$= \frac{1}{M_n} \left(\gamma_n M_n(M_n-1)\right)^\lambda \left(\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')\cdot \right.$$
$$\left. \cdot\left(\sum_{\boldsymbol{y}} \sqrt{\sum_{s_0=1}^A W^n(\boldsymbol{y}|\boldsymbol{x},s_0)\sum_{s_0'=1}^A W^n(\boldsymbol{y}|\boldsymbol{x}',s_0')}\right)^{\frac{1}{\lambda}}\right)^\lambda\right) \tag{140}$$

where in (140) we simplified the outer A with those under the square root. Taking the negative normalized logarithm in (140) we finally get a lower bound on the exponent of a typical code from an ensemble over a finite state channel with initial state s_0 for a given n. Thus (9) can be rewritten as:

$$E_{\text{trc,x}}(R_n,Q^n) = F_x^n(\hat{\lambda}_n,Q^n) - 2\hat{\lambda}_n R_n - \delta_n \tag{141}$$

where

$$F_x^n(\lambda,Q^n) = -\frac{1}{n}\log\left(\sum_{\boldsymbol{x}}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')\cdot \right.$$
$$\left. \cdot\left(\sum_{\boldsymbol{y}} \sqrt{\sum_{s_0=1}^A W^n(\boldsymbol{y}|\boldsymbol{x},s_0)\sum_{s_0'=1}^A W^n(\boldsymbol{y}|\boldsymbol{x}',s_0')}\right)^{\frac{1}{\lambda}}\right)^\lambda\right). \tag{142}$$

We next show that the limit of (142) for n that tends to infinity exists and is finite. We start with the following lemma, which is the equivalent for TRC of [3, Lemma (5.9.1)].

Lemma 2: For any finite-state channel the following holds:

$$F_x^n(\lambda,Q^n) \geq \frac{k}{n}F_k(\lambda,Q^k) + \frac{l}{n}F_l(\lambda,Q^l) \tag{143}$$

where k and l are positive integers and $k+l=n$.

Proof: See Appendix A. ∎

The next lemma is the equivalent of [3, Lemma (5.9.2)] for the typical error exponent case.

Lemma 3: Let us define:

$$F_x^\infty(\lambda,Q) = \sup_n F_x^n(\lambda,Q^n), \tag{144}$$

Q being the limiting distribution of Q^n, which is assumed to exist. For all pairwise-independent code ensembles and all finite state channels for which there exists an n_0 such that for $n > n_0$ the normalized relative entropy (76)

$$\frac{1}{n}D\left(P^n(\boldsymbol{x},\boldsymbol{x}')\|Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}')\right) < \infty \tag{145}$$

and for $\lambda \geq 1$ we have:

$$\lim_{n\to\infty} F_x^n(\lambda,Q^n) = F_x^\infty(\lambda,Q). \tag{146}$$

Furthermore, for $1 \leq \lambda < \infty$ the convergence is uniform in λ and $F_x^\infty(\lambda,Q)$ is uniformly continuous in λ.

Proof: Consider first the case $1 \leq \lambda < \infty$. The boundedness and positiveness of the derivative (76), particularized to finite-state channels, for any finite λ together with the fact that $F_x^n(1,Q^n)$ is finite (see proof of Lemma 4 below and [3, Lemma (5.9.2)]) implies that $F_x^n(\lambda,Q^n)$ is positive and bounded. This fact together with Lemma 2 allows us to apply [3, Lemma (4A.2)], which implies (146). Furthermore, the finiteness of the derivative in λ for each n implies uniform convergence and uniform continuity.

For the case in which $\hat{\lambda}_n \to \infty$, we have that the following limit:

$$\lim_{n\to\infty} F_x^n(\hat{\lambda}_n,Q^n) \tag{147}$$

exists and is either finite or infinite. This follows from the fact that the derivative of $F_x^n(\lambda,Q^n)$ with respect to λ exists for any n and is positive. ∎

An important implication of Lemma 2 is that it guarantees that the limit in (62) exists. This means that Theorem 2 gives a result which is non trivial and in fact practically relevant.

The behavior of $E_{\mathrm{trc,lb}}(R_n, Q^n)$ in the case $\hat{\lambda}_n \to \infty$ can be derived from the general case presented in Section III by substituting the Bhattacharyya bound on the pairwise error probability with:

$$\sum_{\boldsymbol{y}} \sqrt{\sum_{s_0=1}^{A} W^n(\boldsymbol{y}|\boldsymbol{x}, s_0) \sum_{s_0'=1}^{A} W^n(\boldsymbol{y}|\boldsymbol{x}', s_0')}. \quad (148)$$

From the discussion above it follows that the limit (147) is always ∞ for $R > 0$ (the dependency on R_n is hidden in $\hat{\lambda}_n$, which is in turn a function of R_n), since $\hat{\lambda}_n \to \infty$ for a positive rate R only if $R < R_\infty(Q^n)$, while it can be finite in $R = 0$.

B. Lower Bound on $E_{\mathrm{trc,r}}(R_n, Q^n)$ for Finite-State Channels

We recall from Definition 2 that $E_{\mathrm{trc,r}}(R_n, Q^n) = E_{\mathrm{r}}^n(R_n, Q^n) - \iota_n$ where $E_{\mathrm{r}}^n(R_n, Q^n)$ is the random-coding exponent while, under the conditions of Theorem 2, $\iota_n \to 0$. The derivation of a lower bound to $E_{\mathrm{r}}^n(R_n, Q^n)$ for finite-state channels was developed in [3, Sec. 5.9], leading to the random-coding error exponent

$$E_{\mathrm{r}}(R, Q) = \max_{0 \le \rho \le 1} F_0^\infty(\rho, Q) - \rho R \quad (149)$$

where $F_0^\infty(\rho, Q)$ is defined as

$$F_0^\infty(\rho, Q) = \lim_{n \to \infty} F_0^n(\rho, Q^n) \quad (150)$$

and is related to the limiting E_0-function as

$$F_0^\infty(\rho, Q) = \min_{s_0} E_0^\infty(\rho, Q, s_0) \quad (151)$$

with

$$E_0^\infty(\rho, Q, s_0) = \lim_{n \to \infty} -\frac{1}{n} \log \left(\sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)^{\frac{1}{1+\rho}} \right)^{1+\rho} \right). \quad (152)$$

C. Rate R^*

Let us now focus on indecomposable channels[1]. For such channels different definitions of capacity exist. In [3, Sec. 4.6] the upper ($\overline{C}$) and lower ($\underline{C}$) capacities are considered. They differ in that the former includes a maximization over the initial state s_0, while the latter includes a minimization. Interestingly, for indecomposable channels the two definitions coincide, i.e., $\overline{C} = \underline{C} = C$. In the following we show that there exists a rate R^* such that below R^* $E_{\mathrm{trc,x}}(R_n, Q^n)$ is larger than $E_{\mathrm{trc,r}}(R_n, Q^n)$ while the other way round is true for rates $R^* < R < \underline{C}$.

We start by pointing out that $E_{\mathrm{r}}(R, Q)$ is convex and strictly decreasing in R for $R < \underline{C}$, as shown in [3], and the optimal ρ is a decreasing function of R so that for R that goes to $\underline{C}$ the

[1] See [3, Sec. 4.6] for a definition.

optimal ρ gets close to 0 while it takes value 1 in a continuous interval that includes the point $R = 0$. As for $F_{\mathrm{x}}^n(\lambda_n, Q^n)$, we can see that R and λ play the same roles as in $E_{\mathrm{x}}(\lambda, (Q))$ as defined in [3, Sec. 5.7], and thus $F_{\mathrm{x}}^n(\lambda_n, Q^n) - \lambda_n R_n$ is convex and strictly decreasing in R and the optimal λ is also a decreasing function in R. Note, however, that such quantity (which is a bound on the expurgated exponent for finite-state channels) can reach 0 at a rate which is much smaller than the capacity. Now let us call R_{cr} the largest rate at which the maximum of $E_{\mathrm{r}}(R, Q)$ is obtained for $\rho = 1$.

The following lemma is instrumental in proving Theorem 3, where the TRC exponent for all rates is presented.

Lemma 4: For finite-state channels:

$$F_0^n(1, Q^n) - \frac{\log A}{n} \le F_{\mathrm{x}}^n(1, Q^n) \le F_0^n(1, Q^n) + \frac{\log A}{n}. \quad (153)$$

Furthermore, if $F_0^n(1, Q^n)$ is finite $F_{\mathrm{x}}^n(1, Q^n)$ is finite as well.

Proof: See Appendix B. ∎

The following theorem shows that, for finite-state channels, the limits in Theorem 2 exist and, if $\delta_n \to 0$, the result of the theorem is non-trivial, i.e., the typical exponent exists and is strictly positive. Furthermore, it provides the result to the minimization in Theorem 2 for each rate below capacity.

Theorem 3: For all finite-state channels for which $\delta_n \to 0$ the following holds:

$$\liminf_{n \to \infty} \min \{ E_{\mathrm{trc,x}}(R_n, Q^n), E_{\mathrm{trc,r}}(R_n, Q^n) \}$$
$$= \begin{cases} E_{\mathrm{ex}}(2R, Q) + R & R \le R^* \\ E_{\mathrm{r}}(R, Q) & R > R^* \end{cases} \quad (154)$$

where $R^* = \frac{R_{\mathrm{cr}}}{2}$. Furthermore, $E_{\mathrm{ex}}(2R, Q) + R$ is strictly positive for range $R \le R^*$ while $E_{\mathrm{r}}(R, Q)$ is strictly positive in $R^* < R < \underline{C}$ if a maximization over Q is carried out.

Proof: Using Lemma 4, the fact that both $F_{\mathrm{x}}^n(\lambda, Q^n)$ and $F_0^n(\rho, Q^n)$ are increasing in λ and ρ, respectively, and the fact that $\lambda \ge 1$ while $0 < \rho \le 1$, it follows for large n that

$$\max_{\lambda \ge 1} E_{\mathrm{x}}^n(\lambda, Q^n) - \lambda R_n \ge \max_{0 \le \rho \le 1} F_0^n(\rho, Q^n) - \rho R_n \quad (155)$$

when $R \le R_{\mathrm{cr}}$ and that

$$\max_{\lambda \ge 1} E_{\mathrm{x}}^n(\lambda, Q^n) - \lambda R_n < \max_{0 \le \rho \le 1} F_0^n(\rho, Q^n) - \rho R_n \quad (156)$$

for $R_n > R_{\mathrm{cr}}$. Thus, for $R_n \le R_{\mathrm{cr}}$ we obtain that

$$E_{\mathrm{ex}}^n(R_n, Q^n) \ge E_{\mathrm{r}}(R_n, Q^n) \quad (157)$$

while for $R_n > R_{\mathrm{cr}}$

$$E_{\mathrm{ex}}^n(R_n, Q^n) < E_{\mathrm{r}}(R_n, Q^n) \quad \text{for } R_n > R_{\mathrm{cr}}. \quad (158)$$

Finally, since $E_{\mathrm{trc,x}}(R_n, Q^n) = E_{\mathrm{ex}}^n(2R_n, Q^n) + R_n - \delta_n$, for all cases such that $\delta_n \to 0$ and with reference to the statement of Theorem 2:

$$\liminf_{n \to \infty} \min \{ E_{\mathrm{trc,x}}(R_n, Q^n), E_{\mathrm{trc,r}}(R_n, Q^n) \}$$
$$= \begin{cases} \liminf_{n \to \infty} E_{\mathrm{trc,x}}(R_n, Q^n) = E_{\mathrm{ex}}(2R, Q) + R & 0 \le R \le R^* \\ \liminf_{n \to \infty} E_{\mathrm{trc,r}}(R_n, Q^n) = E_{\mathrm{r}}(R, Q) & R > R^* \end{cases}$$
$$(159)$$

which was obtained by equating $E_{\text{ex}}(2R, Q) + R$ and $E_{\text{r}}(R, Q)$, from the definition of R_{cr} and setting $R^* = \frac{R_{\text{cr}}}{2}$. At the right-hand side of (159), the quantity $E_{\text{ex}}(2R, Q) + R$ is positive for rates $R \leq R^*$, while Gallager's exponent $E_{\text{r}}(R, Q)$ can be made positive up to capacity by maximizing over the input distribution, which gives the theorem statement. ∎

D. State Known at the Receiver

In the following we specialize the results for the FSC to the case in which the state of the channel is known at the receiver. We consider a channel in which the state changes at any symbol time according to a Markov chain with a finite number A of states. In particular, we consider an FSC in which the channel state sequence $\boldsymbol{s}$ seen by a transmitted codeword is a deterministic function of the output $\boldsymbol{y}$ and the initial state s_0, i.e., $\boldsymbol{s} = f(\boldsymbol{y}, s_0)$. At a given instant, the channel is modelled as a DMC, the characteristics of which depend on the state at previous instant. This introduces memory in the channel. Channels corresponding to different states are different DMCs, with the only constraint that the input alphabet $\mathcal{X}$ be the same. An example of such family of channels are FSCs for which the DMCs have the same input set but disjoint output sets, that is, the intersection of the output alphabets of any two DMCs is empty. This allows the receiver to know what the state in previous instant was. Knowing the channel state allows the receiver to use maximum-likelihood decoding with the transition probability $W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)$ and to write (140) as

$$P_{\text{e}}(\mathcal{C}_n, s_0) \leq \frac{1}{M_n} (\gamma_n M_n(M_n - 1))^{\lambda} \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) \cdot \right.$$
$$\left. \cdot Q^n(\boldsymbol{x}') \left(\sum_{\boldsymbol{y}} \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}, s_0) W^n(\boldsymbol{y}|\boldsymbol{x}', s_0)} \right)^{\frac{1}{\lambda}} \right)^{\lambda},$$
(160)

where the initial state s_0 is known at the receiver. Similarly to what done in [3, Sec. 5.9] to derive the random coding exponent (RCE), let us introduce the following conditional probability function:

$$W^n(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{x}, s_0) = \begin{cases} W^n(\boldsymbol{y}|\boldsymbol{x}, s_0) & \text{if} \quad \boldsymbol{s} = \boldsymbol{s}(\boldsymbol{y}, s_0) \\ 0 & \text{otherwise.} \end{cases}$$
(161)

Using (161) in (160) we have:

$$P_e(\mathcal{C}_n, s_0)$$
$$\leq \frac{1}{M_n} (\gamma_n M_n(M_n - 1))^{\lambda} \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) \cdot \right.$$
$$\left. \cdot Q^n(\boldsymbol{x}') \left(\sum_{\boldsymbol{s}} \sum_{\boldsymbol{y}} \sqrt{W^n(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{x}, s_0) W^n(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{x}', s_0)} \right)^{\frac{1}{\lambda}} \right)^{\lambda}$$
$$= \frac{1}{M_n} (\gamma_n M_n(M_n - 1))^{\lambda} \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \cdot \right.$$
$$\left. \cdot \left(\sum_{\boldsymbol{s}} \prod_{i=1}^{n} \sum_{y} \sqrt{W^n(y, s_i|x_i, s_{i-1}) W^n(y, s_i|x_i', s_{i-1})} \right)^{\frac{1}{\lambda}} \right)^{\lambda}.$$
(162)

Note that in (162) only the state sequence $\boldsymbol{s}$ corresponding to the initial state s_0 and output sequence $\boldsymbol{y}$ actually leads to a non-zero term in the sum over $\boldsymbol{s}$. Similarly to [3, Sec. 5.9], we express the overall bound as a function of a matrix product. Let:

$$\beta(s, s', x, x') = \sum_{y} \sqrt{W(y, s|x, s') W(y, s|x', s')} \quad (163)$$

where $s, s' \in \{0, 1, \ldots, A - 1\}$ and $x, x' \in \mathcal{X}$. We define the following set of matrices

$$\boldsymbol{A}_{x,x'} = \begin{pmatrix} \beta(0, 0, x, x') & \cdots & \beta(0, A - 1, x, x') \\ \beta(1, 0, x, x') & \cdots & \beta(1, A - 1, x, x') \\ \vdots & \vdots & \vdots \\ \beta(A - 1, 0, x, x') & \cdots & \beta(A - 1, A - 1, x, x') \end{pmatrix}.$$
(164)

Using (163) and (164) in (162) we have:

$$P_e(\mathcal{C}_n, s_0) \leq \frac{1}{M_n} (\gamma_n M_n(M_n - 1))^{\lambda} \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) \cdot \right.$$
$$\left. \cdot Q^n(\boldsymbol{x}') \left(\boldsymbol{e}(s_0)^T \left(\prod_{i=1}^{n} \boldsymbol{A}_{x_i, x_i'} \right) \mathbf{1} \right)^{\frac{1}{\lambda}} \right)^{\lambda} \quad (165)$$

where $\boldsymbol{e}(s_0)$ is an $A \times 1$ vector with a 1 in position s_0 and 0 elsewhere, while $\mathbf{1}$ is a $A \times 1$ vector of all 1's.

Finding a single-letter expression for (165) is challenging due to the presence of the matrix product, but can be evaluated using Monte Carlo. We do this in the following numerical example, in which the two-state model presented in [3, Fig. 5.9.1] is considered. At a given instant, the channel is modelled as a BSC. The channel transition probability at time t is p_0 if the channel was in state 0 at $t-1$, and p_1 otherwise, the probability of state change being q. As in [3, Figure 5.9.2], we assume that the receiver knows the state of the channel at every instant. For this specific channel, $\boldsymbol{A}_{x,x'}$ in (164) can take two possible values, given by

$$\boldsymbol{A}_{0,0} = \boldsymbol{A}_{1,1} = \begin{pmatrix} 1 - q & q \\ q & 1 - q \end{pmatrix}, \quad (166)$$

while:

$$\boldsymbol{A}_{0,1} = \boldsymbol{A}_{1,0}$$
$$= \begin{pmatrix} 2(1 - q)\sqrt{(1 - p_0)p_0} & 2q\sqrt{(1 - p_1)p_1} \\ 2q\sqrt{(1 - p_0)p_0} & 2(1 - q)\sqrt{(1 - p_0)p_0} \end{pmatrix}.$$
(167)

The bound in (154) is plotted against the rate in Fig. 1. The term $E_{\text{r}}(R, Q)$ given in (149) is evaluated in closed form as in [3, Fig. 5.9.2], while $E_{\text{trc,x}}(R_n, Q^n)$, i.e., the the negative normalized logarithm of (165), is evaluated using the Monte Carlo method with 10^6 iterations for a codeword length of $n = 200$. More specifically, due to channel memory, the statistical mean inside square brackets in Eq. (165) cannot be written as a single-letter product, and can only be estimated by randomly generating a large-enough set of pairs of codewords $(\boldsymbol{x}, \boldsymbol{x}')$ of length n. The TRC for the DMC channel derived from this FSC (i.e., setting $q = 1/2$, see [3, Sec. 5.9]) is also shown
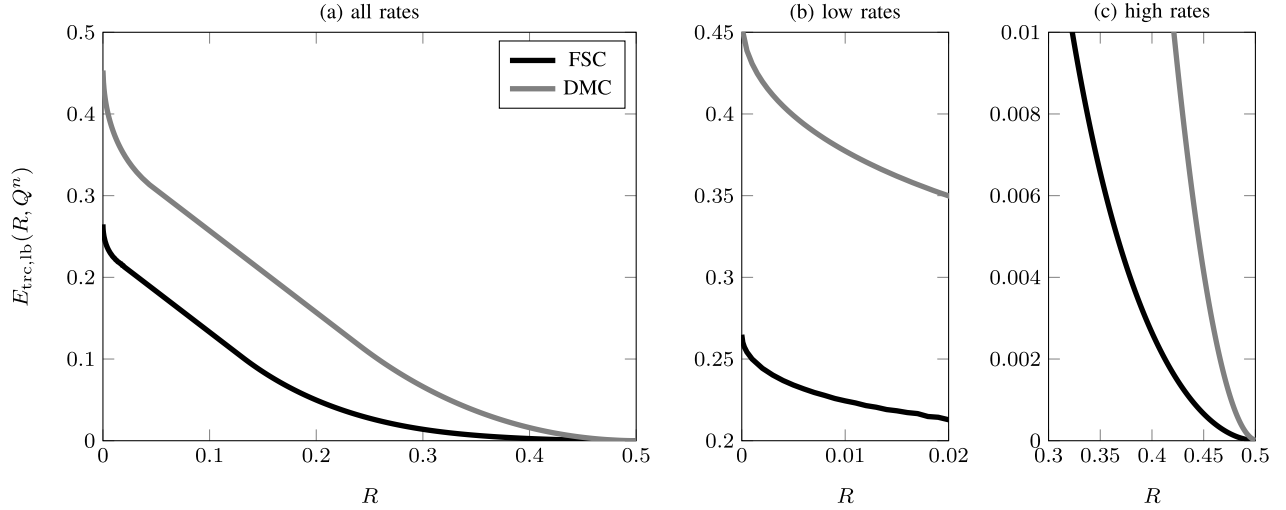
Fig. 1. TRC exponent for the FSC described in [3, Sec. 5.9] with $q = 0.01$, $p_0 = 0.01$ and $p_1 = 0.1$ obtained with monte carlo for $n = 200$ and 10^6 iterations. The TRC over DMC is also shown for comparison. The latter is obtained by setting the state transition probability q to $1/2$.

for comparison. The result is consistent with the behaviour of the RCE presented in [3]. Note that, for this channel, the memory does not decrease the capacity with respect to its i.i.d. counterpart. This can be seen in Fig. 1(c) noting that the smallest rates for which the two curves are zero coincide.

APPENDIX

A. Proof of Lemma 2

Let us split the length-n input sequence $\boldsymbol{x}$ into two sequences of length k and l $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively, i.e., $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$. Let us do the same for the output sequence, i.e., $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$. Let us also impose that the input distribution be $Q^n(\boldsymbol{x}) = Q^n(\boldsymbol{x}_1)Q^n(\boldsymbol{x}_2)$. This choice can be suboptimal in some cases, but does not impact our proof. With this assumptions and from (130) it follows that

$$
e^{-n F_{\mathrm{x}}^n(\lambda, Q^n)}
$$
$$
= \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}) Q^n(\boldsymbol{x}') \cdot \right.
$$
$$
\left. \cdot \left(\sum_{\boldsymbol{y}} \sqrt{ \sum_{s_0=1}^A W^n(\boldsymbol{y}|\boldsymbol{x}, s_0) \sum_{s_0'=1}^A W^n(\boldsymbol{y}|\boldsymbol{x}', s_0') } \right)^{\frac{1}{\lambda}} \right)^{\lambda}
$$
(168)

$$
= \left(\sum_{\boldsymbol{x}_1} \sum_{\boldsymbol{x}_2} \sum_{\boldsymbol{x}_1'} \sum_{\boldsymbol{x}_2'} Q^n(\boldsymbol{x}_1) Q^n(\boldsymbol{x}_2) Q^n(\boldsymbol{x}_1') Q^n(\boldsymbol{x}_2') \cdot \right.
$$
$$
\left. \cdot \left(\sum_{\boldsymbol{y}_1} \sum_{\boldsymbol{y}_2} \sqrt{ \sum_{s_0=1}^A \sum_{s_k=1}^A W^k(\boldsymbol{y}_1, s_k|\boldsymbol{x}_1, s_0) W^l(\boldsymbol{y}_2|\boldsymbol{x}_2, s_k) } \cdot \right. \right.
$$
$$
\left. \left. \cdot \sqrt{ \sum_{s_0'=1}^A \sum_{s_k'=1}^A W^k(\boldsymbol{y}_1, s_k'|\boldsymbol{x}_1', s_0') W^l(\boldsymbol{y}_2|\boldsymbol{x}_2', s_k') } \right)^{\frac{1}{\lambda}} \right)^{\lambda}
$$
(169)

$$
= \left(\sum_{\boldsymbol{x}_1, \boldsymbol{x}_1'} Q^n(\boldsymbol{x}_1) Q^n(\boldsymbol{x}_1') \sum_{\boldsymbol{x}_2, \boldsymbol{x}_2'} Q^n(\boldsymbol{x}_2) Q^n(\boldsymbol{x}_2') \cdot \right.
$$
$$
\left. \cdot \left(\sum_{\boldsymbol{y}_1} \sum_{\boldsymbol{y}_2} \sqrt{ \sum_{s_k=1}^A \sum_{s_k'=1}^A W^l(\boldsymbol{y}_2|\boldsymbol{x}_2, s_k) W^l(\boldsymbol{y}_2|\boldsymbol{x}_2', s_k') } \cdot \right. \right.
$$
$$
\left. \left. \cdot \sqrt{ \sum_{s_0=1}^A \sum_{s_0'=1}^A W^k(\boldsymbol{y}_1, s_k|\boldsymbol{x}_1, s_0) W^k(\boldsymbol{y}_1, s_k'|\boldsymbol{x}_1', s_0') } \right)^{\frac{1}{\lambda}} \right)^{\lambda}
$$
(170)

$$
\leq \left(\sum_{\boldsymbol{x}_1, \boldsymbol{x}_1'} Q^n(\boldsymbol{x}_1) Q^n(\boldsymbol{x}_1') \sum_{\boldsymbol{x}_2, \boldsymbol{x}_2'} Q^n(\boldsymbol{x}_2) Q^n(\boldsymbol{x}_2') \cdot \right.
$$
$$
\left. \cdot \left(\sum_{\boldsymbol{y}_1} \sum_{\boldsymbol{y}_2} \sqrt{ \sum_{s_k=1}^A \sum_{s_k'=1}^A W^l(\boldsymbol{y}_2|\boldsymbol{x}_2, s_k) W^l(\boldsymbol{y}_2|\boldsymbol{x}_2', s_k') } \cdot \right. \right.
$$
$$
\left. \left. \cdot \sqrt{ \sum_{s_0=1}^A \sum_{s_0'=1}^A W^k(\boldsymbol{y}_1|\boldsymbol{x}_1, s_0) W^k(\boldsymbol{y}_1|\boldsymbol{x}_1', s_0') } \right)^{\frac{1}{\lambda}} \right)^{\lambda}
$$
(171)

$$
= \left(\sum_{\boldsymbol{x}_1, \boldsymbol{x}_1'} Q^n(\boldsymbol{x}_1) Q^n(\boldsymbol{x}_1') \right.
$$
$$
\left(\sum_{\boldsymbol{y}_1} \sqrt{ \sum_{s_0=1}^A \sum_{s_0'=1}^A W^k(\boldsymbol{y}_1|\boldsymbol{x}_1, s_0) W^k(\boldsymbol{y}_1|\boldsymbol{x}_1', s_0') } \right)^{\frac{1}{\lambda}} \right)^{\lambda} \cdot
$$
$$
\cdot \left(\sum_{\boldsymbol{x}_2, \boldsymbol{x}_2'} Q^n(\boldsymbol{x}_2) Q^n(\boldsymbol{x}_2') \right.
$$
$$
\left(\sum_{\boldsymbol{y}_2} \sqrt{ \sum_{s_k=1}^A \sum_{s_k'=1}^A W^l(\boldsymbol{y}_2|\boldsymbol{x}_2, s_k) W^l(\boldsymbol{y}_2|\boldsymbol{x}_2', s_k') } \right)^{\frac{1}{\lambda}} \right)^{\lambda}
$$
(172)

$$
= e^{-k F_k(\lambda, Q^k)} e^{-l F_l(\lambda, Q^l)}
$$
(173)

where in (169) we applied (131), in (170) we reorganized (169), while in (171) we used the fact that

$$
W^k(\boldsymbol{y}_1, s_k|\boldsymbol{x}_1, s_0)W^k(\boldsymbol{y}_1, s_k'|\boldsymbol{x}_1', s_0')
$$

$$
\leq \sum_{s_k, s_k'} W^k(\boldsymbol{y}_1, s_k|\boldsymbol{x}_1, s_0)W^k(\boldsymbol{y}_1, s_k'|\boldsymbol{x}_1', s_0') \tag{174}
$$

$$
= W^k(\boldsymbol{y}_1|\boldsymbol{x}_1, s_0)W^k(\boldsymbol{y}_1|\boldsymbol{x}_1', s_0') \tag{175}
$$

which concludes the proof.

B. Proof of Lemma 4

Let us evaluate $F_x^n(\lambda, Q^n)$ at $\lambda = 1$. From (142) we obtain that

$$
e^{-nF_x^n(1,Q^n)}
$$

$$
= \sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}') \cdot
$$

$$
\cdot \sum_{\boldsymbol{y}} \sqrt{\sum_{s_0=1}^{A} W^n(\boldsymbol{y}|\boldsymbol{x}, s_0) \sum_{s_0'=1}^{A} W^n(\boldsymbol{y}|\boldsymbol{x}', s_0')} \tag{176}
$$

$$
= \sum_{\boldsymbol{y}} \sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x})Q^n(\boldsymbol{x}') \cdot
$$

$$
\cdot \sqrt{\sum_{s_0=1}^{A} W^n(\boldsymbol{y}|\boldsymbol{x}, s_0) \sum_{s_0'=1}^{A} W^n(\boldsymbol{y}|\boldsymbol{x}', s_0')} \tag{177}
$$

$$
= \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sqrt{\sum_{s_0=1}^{A} W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)} \right)^2 \tag{178}
$$

$$
\leq \sum_{\boldsymbol{y}} \left(\sum_{s_0=1}^{A} \frac{A}{A} \sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)} \right)^2 \tag{179}
$$

$$
= A^2 \sum_{\boldsymbol{y}} \left(\sum_{s_0=1}^{A} \frac{1}{A} \sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)} \right)^2 \tag{180}
$$

$$
\leq A^2 \sum_{s_0=1}^{A} \frac{1}{A} \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)} \right)^2 \tag{181}
$$

$$
= A \sum_{s_0=1}^{A} \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)} \right)^2 \tag{182}
$$

$$
\leq A^2 \max_{s_0} \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)} \right)^2 \tag{183}
$$

$$
= Ae^{-nF_0^n(1,Q^n)}. \tag{184}
$$

On the other hand,

$$
e^{-nF_x^n(1,Q^n)} = \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sqrt{\sum_{s_0=1}^{A} W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)} \right)^2 \tag{185}
$$

$$
\geq \max_{s_0} \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q^n(\boldsymbol{x}) \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}, s_0)} \right)^2 \tag{186}
$$

$$
= \frac{1}{A} e^{-nF_0^n(1,Q^n)} \tag{187}
$$

and, since A is finite, we have that

$$
F_0^n(1, Q^n) - \frac{\log A}{n} \leq F_x^n(1, Q^n) \leq F_0^n(1, Q^n) + \frac{\log A}{n}. \tag{188}
$$

Using (188) and the fact that A is finite, $F_x^n(1, Q^n)$ coincides asymptotically with $F_0^n(1, Q^n)$. Finally, since the chains of inequality above hold for any n and A is finite, if $F_0^n(1, Q^n)$ is finite $F_x^n(1, Q^n)$ is also finite.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, Oct. 1948.

[2] R. M. Fano, *Transmission of Information: A Statistical Theory of Communication*. Cambridge, MA, USA: Massachusetts Institute of Technology Press, 1961.

[3] R. G. Gallager, *Information Theory and Reliable Communication*. Hoboken, NJ, USA: Wiley, 1968.

[4] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding in discrete memoryless channels I-II," *Inf. Control*, vol. 10, pp. 65–103, 522–552, 1967.

[5] I. Csiszár, J. Körner, and K. Marton, "A new look at exponential error bounds for memoryless channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Ithaca, NY, USA, Oct. 1977.

[6] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 1, pp. 5–12, Jan. 1981.

[7] A. Somekh-Baruch, J. Scarlett, and A. Guillén i Fàbregas, "Generalized random Gilbert-Varshamov codes," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3452–3469, Jun. 2019.

[8] A. Barg and G. D. Forney, "Random codes: Minimum distances and error exponents," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2568–2573, Sep. 2002.

[9] A. Nazari, A. Anastasopoulos, and S. S. Pradhan, "Error exponent for multiple-access channels: Lower bounds," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5095–5115, Sep. 2014.

[10] N. Merhav, "Error exponents of typical random codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6223–6235, Sep. 2018.

[11] N. Merhav, "Error exponents of typical random trellis codes," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2067–2077, Apr. 2020.

[12] N. Merhav, "Error exponents of typical random codes for the colored Gaussian channel," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8164–8179, Dec. 2019.

[13] R. Tamir, N. Merhav, N. Weinberger, and A. Guillén i Fàbregas, "Large deviations behavior of the logarithmic error probability of random codes," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 6635–6659, Nov. 2020.

[14] L. V. Truong, G. Cocco, J. Font-Segura, and A. Guillén i Fàbregas, "Concentration properties of random codes," 2022, *arXiv:2203.07853*.

[15] R. Tamir and N. Merhav, "Universal decoding for the typical random code and for the expurgated code," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2156–2168, Apr. 2022.

[16] N. Merhav, "A Lagrange–dual lower bound to the error exponent of the typical random code," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3456–3464, Jun. 2020.

[17] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliott channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 6, pp. 1277–1290, Nov. 1989.

[18] N. Merhav, "On zero-rate error exponents of finite-state channels with input-dependent states," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 741–750, Feb. 2015.

[19] D. Blackwell, L. Breiman, and A. J. Thomasian, "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Statist.*, vol. 29, no. 4, pp. 1209–1220, 1958.

[20] D. Blackwell, "Exponential error bounds for finite-state channels," in *Proc. 4th Berkeley Symp. Math. Stat. Prob.*, vol. 1, Berkeley, California, Jun. 1960, pp. 57–63.

[21] H. Yudkin, "On the exponential error bound and capacity for finite-state channels," in *Proc. Int. Symp. Info. Theory*, San Remo, Italy, Sep. 1967, pp. 1386–1390.

[22] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[23] N. Merhav, "The generalized stochastic likelihood decoder: Random coding and expurgated bounds," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5039–5051, Aug. 2017.

[24] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 1, pp. 3–18, Jan. 1965.

[25] S. I. Resnick, *A Probability Path*. Boston, MA, USA: Birkhäuser, 2005.

[26] R. Durrett, *Probability: Theory Examples*, 5th ed. Cambridge, U.K.: Cambridge Univ. Press, Apr. 2019.

[27] J. Font-Segura, A. Martinez, and A. Guillén i Fàbregas, "Asymptotics of the error probability in quasi-static binary symmetric channels," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2493–2497.

[28] J. Scarlett, A. Guillén i Fàbregas, A. Somekh-Baruch, and A. Martinez, "Information-theoretic foundations of mismatched decoding," *Found. Trends Commun. Inf. Theory*, vol. 17, nos. 2–3, pp. 149–401, 2020.

[29] J. Scarlett, L. Peng, N. Merhav, A. Martinez, and A. Guillén i Fàbregas, "Expurgated random-coding ensembles: Exponents, refinements, and connections," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4449–4462, Aug. 2014.

[30] G. Poltyrev, "Random coding bounds for discrete memoryless channels," *Probl. Peredachi Inf.*, vol. 18, no. 1, pp. 12–26, 1982.

[31] J. Scarlett, A. Martinez, and A. I. Guillén i Fàbregas, "Ensemble-tight error exponents for mismatched decoders," in *Proc. Annu. Allerton Conf. Commun., Control, Comput.*, Allerton, IL, USA, Oct. 2012, pp. 1951–1958.

[32] C. Stone, "On local and ratio limit theorems," in *Proc. Berkeley Symp. Math. Stat. Prob.*, 1965, pp. 217–224.

Giuseppe Cocco (Senior Member, IEEE) received the M.S. degree (Hons.) in telecommunications engineering from Università di Pisa, Pisa, Italy, in 2008, and the Ph.D. degree in telecommunications engineering from the Universitat Politècnica de Catalonia (UPC), Barcelona, Spain, in 2013.

In 2009, he was awarded the Networking/Partnering Initiative grant (Ph.D. co-funding) by the European Space Agency (ESA). His Ph.D. studies were co-funded by ESA and the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Barcelona. From 2010 to 2011, he spent one year as Visiting Ph.D. Student at the European Space Research and Technology Centre (ESTEC), ESA, Noordwijk, The Netherlands, working on the application of network coding in satellite communication systems. From 2014 to 2017, he was at the Institute of Communications and Navigation, German Aerospace Center (DLR), Weßling, Germany, where he was Senior Researcher and Project Manager. In 2017, he was awarded a Marie Skłodowska Curie Individual fellowship grant by the European Commission and from 2017 to 2019 he was at the Laboratory of Intelligent Systems (robotics) and the LTS4 Signal Processing Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. In 2019, he was awarded the Beatriu de Pinós fellowship by the Catalan Government and joined the Information Theory and Coding group at the Universitat Pompeu Fabra (UPF), Barcelona. In 2022, he was awarded the Ramon y Cajal Fellowship by the Spanish Ministry of Research. His research interests include information theory, satellite communications, video streaming, interference management, and network coding.

Albert Guillén i Fàbregas (Fellow, IEEE) received the degree in telecommunications engineering and the degree in electronics engineering from Universitat Politècnica de Catalunya and Politecnico di Torino, respectively, in 1999 and the Ph.D. degree in communication systems from École Polytechnique Fédérale de Lausanne (EPFL) in 2004.

In 2020, he returned to a full-time faculty position at the Department of Engineering, University of Cambridge, where he was a full-time faculty and a fellow of Trinity Hall from 2007 to 2012. Since 2011, he has been an ICREA Research Professor with Universitat Pompeu Fabra (currently on leave). He has held appointments at the New Jersey Institute of Technology, Telecom Italia, European Space Agency (ESA), Institut Eurécom, University of South Australia, Universitat Pompeu Fabra, University of Cambridge, as well as visiting appointments at EPFL, École Nationale des Télécommunications (Paris), Universitat Pompeu Fabra, University of South Australia, Centrum Wiskunde and Informatica and Texas A&M University in Qatar. His specific research interests include information theory, communication theory, coding theory, and statistical inference.

Dr. Guillén i Fàbregas is a member of the Young Academy of Europe, and received the Starting and Consolidator Grants from the European Research Council, the Young Authors Award of the 2004 European Signal Processing Conference, the 2004 Best Doctoral Thesis Award from the Spanish Institution of Telecommunications Engineers, and a Research Fellowship of the Spanish Government to join ESA. He is a fellow of the Institute for Mathematics and its Applications (IMA). Since 2013, he has been an Editor of the *Foundations and Trends in Communications and Information Theory* (Now Publishers) and was an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY from 2013 to 2020 and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2007 to 2011.

Josep Font-Segura (Member, IEEE) was born in Barcelona, Spain, in 1984. He received the Ph.D. degree in electrical engineering from Universitat Politècnica de Catalunya, Barcelona, in 2014.

From 2008 to 2009, he was a Research Assistant at the Information Signal Processing Group, Columbia University, New York. Since 2014, he has been a Research Associate at Universitat Pompeu Fabra, Barcelona, with the Information Theory and Coding Group. His research interests include information theory, mathematical physics, digital communications, and signal processing.

Dr. Font-Segura was a member of the Organizing Committee of the IEEE International Symposium on Information Theory in 2016. He was a recipient of the "la Caixa" Savings Bank Fellowship for graduate studies in the USA, the Best Paper Award at the IEEE International Conference on Communications, and the Juan de la Cierva Postdoctoral Fellowship from the Spanish Government. He is currently the Vice-Chair of the Spanish Chapter of the IEEE Information Theory Society. He also contributed as reviewer in several journal and conference publications of the IEEE society.