Expurgated Random-Coding Ensembles: Exponents, Refinements and Connections

Jonathan Scarlett, Li Peng, Neri Merhav, Alfonso Martinez and Albert Guillén i Fàbregas

Abstract

This paper studies expurgated random-coding bounds and exponents for channel coding with a given (possibly suboptimal) decoding rule. Variations of Gallager's analysis are presented, yielding several asymptotic and non-asymptotic bounds on the error probability for an arbitrary codeword distribution. A simple non-asymptotic bound is shown to attain an exponent of Csiszár and Körner under constant-composition coding. Using Lagrange duality, this exponent is expressed in several forms, one of which is shown to permit a direct derivation via cost-constrained coding which extends to infinite and continuous alphabets. The method of type class enumeration is studied, and it is shown that this approach can yield improved exponents and better tightness guarantees for some codeword distributions. A generalization of this approach is shown to provide a multi-letter exponent which extends immediately to channels with memory. Finally, a refined analysis expurgated i.i.d. random coding is shown to yield a $O\big(\frac{1}{\sqrt{n}}\big)$ prefactor, thus improving on the standard $O(1)$ prefactor. Moreover, the implied constant is explicitly characterized.

Index Terms

Expurgated error exponents, reliability function, random coding, mismatched decoding, maximum-likelihood decoding, type class enumeration

I. Introduction

Achievable performance bounds for channel coding are typically obtained by analyzing the average error probability of an ensemble of codebooks with independently generated codewords. For memoryless channels, random codes with independent and identically distributed (i.i.d.) symbols achieve the channel capacity [1], characterize the error exponent of the best code at sufficiently high rates [2, Ch. 5], and provide tight bounds on the finite-length performance [3].

J. Scarlett and L. Peng are with the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mails: jmscarlett@gmail.com, lp327@cam.ac.uk). N. Merhav is with the Department of Electrical Engineering, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel. (e-mail: merhav@ee.technion.ac.il). A. Martinez is with the Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08018 Barcelona, Spain (e-mail: alfonso.martinez@ieee.org). A. Guillén i Fàbregas is with the Institució Catalana de Recerca i Estudis Avançats (ICREA), the Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08018 Barcelona, Spain, and also with the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mail: guillen@ieee.org).

This work has been funded in part by the European Research Council under ERC grant agreement 259663, by the European Union's 7th Framework Programme (PEOPLE-2011-CIG) under grant agreement 303633 and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150 and TEC2012-38800-C03-03. The work of N. Merhav was partially supported by the Israel Science Foundation (ISF), grant no. 412/12.

This is the extended version of a paper which was accepted to *IEEE Transactions on Information Theory* (April 2014). A shortened version was presented at the 2014 International Zurich Seminar on Communications.

At low rates, the error probability of the best code in the random-coding ensemble can be significantly smaller than the average. In such cases, better performance bounds are obtained by considering an ensemble in which a subset of the randomly generated codewords are expurgated from the codebook. In particular, the error exponents resulting from such techniques are generally higher than the random-coding error exponent at low rates. Existing works exploring such techniques include those of Gallager [2, Sec. 5.7], Csiszár-Körner-Marton [4], [5, Ex. 10.18] and Csiszár-Körner [6]. The advantages of Gallager's approach include its simplicity and the fact that the analysis is not restricted to finite alphabets. On the other hand, as we will see in Section III, the exponents of [4]–[6] can improve on that of Gallager for a given input distribution or decoding rule.

In this paper, we provide techniques that attain the best of each of the above approaches. Using variations of Gallager's analysis, we obtain several asymptotic and non-asymptotic bounds for an arbitrary codeword distribution. Using these bounds, we provide derivations of both new and existing expurgated exponents, each yielding various advantages such as simplicity, generality, and guarantees of exponential tightness. We explore the method of type class enumeration (e.g. see [7]–[9]) for both discrete and continuous channels, and show that it can yield improved exponents and tightness guarantees, as well as providing a multi-letter exponent which extends immediately to channels with memory.

A. System Setup

The input and output alphabets are denoted by $\mathcal{X}$ and $\mathcal{Y}$ respectively. The channel is assumed to be memoryless, yielding an n-letter transition law given by $W^n(\boldsymbol{y}|\boldsymbol{x}) \triangleq \prod_{i=1}^n W(y_i|x_i)$ for some conditional distribution $W(y|x)$. In the case that both $\mathcal{X}$ and $\mathcal{Y}$ are finite, the channel is a discrete memoryless channel (DMC), but we do not assume this to be the case in general. The encoder takes as input a message m equiprobable on the set $\{1,\dots,M\}$, and transmits the corresponding codeword $\boldsymbol{x}^{(m)}$ from a codebook $\mathcal{C} = \{\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(M)}\}$. The decoder receives the vector $\boldsymbol{y}$ at the output of the channel, and forms the estimate

$$\hat{m} = \operatorname*{arg\,max}_{j\in\{1,\dots,M\}} q^n(\boldsymbol{x}^{(j)},\boldsymbol{y}), \tag{1}$$

where $q^n(\boldsymbol{x},\boldsymbol{y}) \triangleq \prod_{i=1}^n q(x_i,y_i)$, and $q(x,y)$ is a non-negative function called the *decoding metric*. An error is said to have occurred if $\hat{m} \neq m$, and we assume that ties are broken as errors. We let $p_{e,m}(\mathcal{C})$ be the error probability induced by $\mathcal{C}$ given a particular message m, and we denote the maximal error probability by $p_e(\mathcal{C}) \triangleq \max_m p_{e,m}(\mathcal{C})$.

When $q(x,y) = W(y|x)$, (1) is the optimal maximum-likelihood (ML) decoding rule. For other decoding metrics, this setting is that of *mismatched decoding* [10]–[13], which is relevant when ML decoding is not feasible, e.g. due to channel uncertainty or implementation constraints.

Throughout the paper, we consider channels with both constrained and unconstrained inputs. In the former setting, each codeword $\boldsymbol{x}$ must satisfy a constraint of the form

$$\frac{1}{n}\sum_{i=1}^n c(x_i) \leq \Gamma, \tag{2}$$

where $c(\cdot)$ is referred to as a cost function, and Γ is a constant. Except where stated otherwise, it will be assumed that the input is unconstrained, which corresponds to $\Gamma = \infty$.

For a given rate R, an error exponent $E(R)$ is said to be achievable if there exists a sequence of codebooks $\mathcal{C}_n$ of length n and rate R whose error probability $p_e(\mathcal{C}_n)$ satisfies

$$\liminf_{n\to\infty} -\frac{1}{n}\log p_e(\mathcal{C}_n) \geq E(R). \tag{3}$$

We focus on the maximal error probability rather than the average error probability, but the two are equivalent for the purposes of studying error exponents.

B. Previous Work

Considering ML decoding, Gallager [2, Ch. 5] studied an ensemble in which $2M-1$ codewords are generated at random, and a subset of M codewords forms the codebook. Roughly speaking, the codewords which are kept are those which have the lowest error probability among the original codewords. A different approach was taken by Csiszár, Körner and Marton [4] (see also [5, Ex. 10.18]), who began by proving the existence of a collection of constant-composition codewords such that any two codewords have a joint empirical distribution satisfying certain properties. By analyzing this collection of codewords using the method of types, an error exponent was obtained which coincides with that of Gallager after the optimization of the input distribution. An exponent for mismatched decoding was derived by Csiszár and Körner [6], and was shown to coincide with that of [4] when particularized to the case of ML decoding.

As stated in the introduction, the exponents of [4], [6] can in fact improve on that of Gallager for a given input distribution. However, the proofs rely heavily on techniques which are valid only when the input and output alphabets are finite. In particular, [4] uses the type packing lemma [5, Ch. 10], and [6] uses a combinatorial graph decomposition lemma. For other related works, see [14]–[17].

Overviews of the mismatched decoding problem can be found in [10]–[13]. Most of the literature has focused on achievable rates, whereas this paper is concerned with the performance at low rates. The mismatched decoding paper most relevant to this one is [13], which studies random-coding error exponents for various non-expurgated ensembles.

C. Contributions

The main contributions of this paper are as follows:

- In Section II, we present variations of Gallager's analysis which yield several asymptotic and non-asymptotic bounds on the error probability. In particular, we consider the use of a logarithmic function in the expurgation argument in place of the power function used by Gallager [2, Sec. 7.3].
- In Section III, we present an overview of various expurgated exponents and the connections between them. Using the method of Lagrange duality [18], we relate the exponents given in [2], [4], [6]. Generalizations of the exponents in [2], [4] to the setting of mismatched decoding are given, and an alternative form of the exponent in [6] is given which extends readily to channels with infinite or continuous alphabets.
- In Section IV, we present several methods for deriving both new and existing exponents:
  - In Section IV-A, we present simple techniques for deriving exponents using a non-asymptotic bound from Section II. Applying constant-composition coding and the method of types recovers the exponent in [6], thus

providing a simple and concise proof. Furthermore, applying cost-constrained coding with multiple auxiliary costs [13] recovers the generalization of this exponent to more general alphabets.

  - In Section IV-B, we study the method of type class enumeration (e.g. see [7]–[9]), which is shown to yield better exponents than the simpler approach for some codeword distributions, as well as better guarantees of exponential tightness.
  - In Section IV-C, we extend the type class enumeration analysis to allow for infinite and continuous alphabets. This is not only of interest in itself, but also yields a multi-letter exponent which can be directly applied to channels with memory and more general decoding metrics.

- In Section V, we present a refined derivation of Gallager's exponent for i.i.d. random coding (and its generalization to mismatched decoding) with a $O\big(\frac{1}{\sqrt{n}}\big)$ prefactor, thus improving on the original $O(1)$ prefactor. Similar improvements for the non-expurgated random-coding error exponent have recently been obtained by Altuğ and Wagner [19] (see also [20]).

D. Notation

We use bold symbols for vectors (e.g. $\boldsymbol{x}$), and denote the corresponding i-th entry using a subscript (e.g. x_i).

The set of all probability distributions on an alphabet, say $\mathcal{X}$, is denoted by $\mathcal{P}(\mathcal{X})$, and the set of all empirical distributions on a vector in $\mathcal{X}^n$ (i.e. types [5, Ch. 2]) is denoted by $\mathcal{P}_n(\mathcal{X})$. For a given type $Q \in \mathcal{P}_n(\mathcal{X})$, the type class $T^n(Q)$ is defined to be the set of all sequences in $\mathcal{X}^n$ with type Q.

The probability of an event is denoted by $\mathbb{P}[\cdot]$, and the symbol $\sim$ means "distributed as". The marginals of a joint distribution $P_{XY}(x,y)$ are denoted by $P_X(x)$ and $P_Y(y)$. We write $P_X = \widetilde{P}_X$ to denote element-wise equality between two probability distributions on the same alphabet. Expectation with respect to a joint distribution $P_{XY}(x,y)$ is denoted by $\mathbb{E}_P[\cdot]$, or simply $\mathbb{E}[\cdot]$ when the associated probability distribution is understood from the context. Similarly, the mutual information with respect to P_{XY} is written as $I_P(X;Y)$, or simply $I(X;Y)$ when the distribution is understood from the context. Given a distribution $Q(x)$ and conditional distribution $W(y|x)$, we write $Q \times W$ to denote the joint distribution defined by $Q(x)W(y|x)$.

For two positive sequences f_n and g_n, we write $f_n \doteq g_n$ if $\lim_{n\to\infty} \frac{1}{n} \log \frac{f_n}{g_n} = 0$, and we write $f_n \dot{\leq} g_n$ if $\limsup_{n\to\infty} \frac{1}{n} \log \frac{f_n}{g_n} \leq 0$, and analogously for $\dot{\geq}$. We write $f_n = O(g_n)$ if $|f_n| \leq c|g_n|$ for some c and sufficiently large n. All logarithms have base e, and all rates are in units of nats except in the examples, where bits are used. We define $[c]^+ = \max\{0, c\}$, and denote the indicator function by $\mathbb{1}\{\cdot\}$.

II. Expurgated Bounds

In this section, we present a number of variations of Gallager's bounds and techniques which will provide the starting points of the derivations of the exponents in Section IV. We let $P_{\boldsymbol{X}}$ denote a codeword distribution, and we define the random variables $(\boldsymbol{X}, \boldsymbol{Y}, \overline{\boldsymbol{X}})$ distributed according to

$$(\boldsymbol{X}, \boldsymbol{Y}, \overline{\boldsymbol{X}}) \sim P_{\boldsymbol{X}}(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}) P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}). \tag{4}$$

In the case that a cost constraint of the form (2) is present, we assume that $P_{\boldsymbol{X}}$ is chosen such that $\boldsymbol{X}$ satisfies the constraint with probability one.

We let $\mathsf{C} = \{\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(M')}\}$ be a random codebook of size M' with each codeword independently generated according to $P_{\boldsymbol{X}}$. The symbol $\mathcal{C}$ is used to denote a fixed expurgated codebook containing $M \leq M'$ codewords.

We begin with the following straightforward generalization of [2, Lemma, p. 151].

Lemma 1. *Fix a function $f : [0,1] \to \mathbb{R}$ and a codeword distribution $P_{\boldsymbol{X}}$ such that $f(p_{e,m}(\mathsf{C}))$ is non-negative for all m with probability one. For any $\eta > 0$, there exists a codebook $\mathcal{C}$ of size M such that $M' \frac{\eta}{1+\eta} < M \leq M'$ and*

$$f\big(p_{e,m}(\mathcal{C})\big) \leq (1+\eta)\mathbb{E}\big[f(p_{e,m}(\mathsf{C}))\big] \tag{5}$$

for $m = 1, \ldots, M$.

Proof: The proof is identical to [2, Lemma, p. 151], with the assumption of $f(p_{e,m}(\mathsf{C}))$ being non-negative ensuring the validity of Markov's inequality. ∎

While Lemma 1 is valid for any function $f(\cdot)$, it is primarily of interest when $f(\cdot)$ is monotonically increasing, so that (5) can be inverted in order to obtain an upper bound on $p_{e,m}(\mathcal{C})$. Gallager [2] presented the lemma with the choices $\eta = 1$ and $f(\cdot) = (\cdot)^{1/\rho}$, where $\rho > 0$, thus proving the existence of a codebook $\mathcal{C}$ of size M such that

$$p_e(\mathcal{C}) \leq \Big(2\mathbb{E}\big[p_{e,m}(\mathsf{C})^{1/\rho}\big]\Big)^{\rho}, \tag{6}$$

where C contains $M' = 2M - 1$ codewords. In the following theorem, we provide non-asymptotic bounds on the error probability which follow in a straightforward fashion from (6). The proof alters Gallager's arguments for the purpose of better characterizing the non-asymptotic performance, and also for dealing with suboptimal decoding rules.

Theorem 1. *For any pair (n, M), codeword distribution $P_{\boldsymbol{X}}$, and parameters $\rho \geq 1$ and $s \geq 0$, there exists a codebook $\mathcal{C}_n$ with M codewords of length n whose maximal error probability satisfies*

$$p_e(\mathcal{C}_n) \leq \mathrm{rcux}_{\rho}(n, M) \leq \mathrm{rcux}_{\rho,s}(n, M) \tag{7}$$

where

$$\mathrm{rcux}_{\rho}(n, M) \triangleq \left(4(M-1)\mathbb{E}\left[\mathbb{P}\Big[q^n(\overline{\boldsymbol{X}}, \boldsymbol{Y}) \geq q^n(\boldsymbol{X}, \boldsymbol{Y}) \,\Big|\, \boldsymbol{X}, \overline{\boldsymbol{X}}\Big]^{1/\rho}\right]\right)^{\rho} \tag{8}$$

$$\mathrm{rcux}_{\rho,s}(n, M) \triangleq \left(4(M-1)\mathbb{E}\left[\mathbb{E}\left[\left(\frac{q^n(\overline{\boldsymbol{X}}, \boldsymbol{Y})}{q^n(\boldsymbol{X}, \boldsymbol{Y})}\right)^{s} \,\middle|\, \boldsymbol{X}, \overline{\boldsymbol{X}}\right]^{1/\rho}\right]\right)^{\rho}. \tag{9}$$

Proof: We obtain (8) from (6) by weakening the expectation as follows:

$$\mathbb{E}\big[p_{e,m}(\mathsf{C})^{1/\rho}\big] \leq \mathbb{E}\left[\Big(\sum_{\overline{m} \neq m} \mathbb{P}\Big[q^n(\boldsymbol{X}^{(\overline{m})}, \boldsymbol{Y}) \geq q^n(\boldsymbol{X}^{(m)}, \boldsymbol{Y}) \,\Big|\, \boldsymbol{X}^{(m)}, \boldsymbol{X}^{(\overline{m})}\Big]\Big)^{1/\rho}\right] \tag{10}$$

$$\leq \mathbb{E}\left[2(M-1)\mathbb{P}\Big[q^n(\overline{\boldsymbol{X}}, \boldsymbol{Y}) \geq q^n(\boldsymbol{X}, \boldsymbol{Y}) \,\Big|\, \boldsymbol{X}, \overline{\boldsymbol{X}}\Big]^{1/\rho}\right], \tag{11}$$

where (10) follows from the union bound, and (11) follows using $M' = 2M - 1$ along with the inequality

$$\Big(\sum_i a_i\Big)^{1/\rho} \leq \sum_i a_i^{1/\rho}, \tag{12}$$

which holds for any $\rho \geq 1$. We obtain (9) by applying Markov's inequality to the inner probability in (8). ∎

Following the terminology of Polyanskiy *et al.* [3], we refer to the bounds in (8)–(9) as *expurgated random-coding union* (RCUX) bounds. These bounds are computable for sufficiently symmetric setups, and are thus of independent interest for characterizing the finite-length performance [3]. It should be noted that both rcux_ρ and $\mathrm{rcux}_{\rho,s}$ extend immediately to channels with memory and general decoding metrics (not necessarily single-letter).

The bound $\mathrm{rcux}_{\rho,s}$ was presented by Gallager [2] under ML decoding with $s = \frac{1}{2}$. For the random-coding ensembles we consider, it will be seen that this choice of s is optimal for ML decoding, at least in terms of the error exponent. However, for mismatched decoding it is important to allow for an arbitrary choice of $s \geq 0$.

The following theorem gives an asymptotic bound which follows by using Lemma 1 with a choice of $f(\cdot)$ which differs from that of Gallager.

Theorem 2. *Consider a sequence of codebooks C_n containing $M_n' = e^{nR}$ codewords which are generated independently according to $P_{\boldsymbol{X}}$. Suppose that there exists a non-negative sequence $E(n)$ growing subexponentially in n (i.e. $E(n) \doteq 1$) such that*

$$\mathbb{P}\big[q^n(\overline{\boldsymbol{x}}, \boldsymbol{Y}) \geq q^n(\boldsymbol{x}, \boldsymbol{Y}) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \geq e^{-E(n)} \tag{13}$$

for all $\boldsymbol{x}$ and $\overline{\boldsymbol{x}}$ on the support of $P_{\boldsymbol{X}}$. Then there exists a sequence of codebooks $\mathcal{C}_n$ with M_n codewords such that

$$\lim_{n\to\infty} \frac{1}{n} \log M_n = R \tag{14}$$

and

$$p_e(\mathcal{C}_n) \dot{\leq} \exp\Big(\mathbb{E}[\log p_{e,m}(\mathsf{C}_n)]\Big) \tag{15}$$

$$\leq \exp\Big(\rho\, \mathbb{E}\Big[\log \mathbb{E}\big[p_{e,m}(\mathsf{C}_n)^{1/\rho} \,\big|\, \boldsymbol{X}^{(m)}\big]\Big]\Big), \tag{16}$$

where (16) *holds for any $\rho > 0$.*

Proof: The error probability associated with the transmitted codeword $\boldsymbol{x}$ is lower bounded by the left-hand side of (13), where $\overline{\boldsymbol{x}}$ is any incorrect codeword. The assumption in (13) thus implies that the function $f(p_{e,m}(\mathsf{C})) = E(n) + \log p_{e,m}(\mathsf{C})$ is non-negative for $m = 1, \ldots, M$. Applying Lemma 1, we obtain that for each n and any $\eta_n > 0$ there exists a codebook $\mathcal{C}_n$ of size $M_n = e^{nR\frac{\eta_n}{1+\eta_n}}$ such that

$$E(n) + \log p_e(\mathcal{C}_n) \leq (1+\eta_n)\big(E(n) + \mathbb{E}[\log p_{e,m}(\mathsf{C}_n)]\big). \tag{17}$$

Since $\log \alpha \leq 0$ for $\alpha \in (0, 1]$, it follows that

$$\log p_e(\mathcal{C}_n) \leq \eta_n E(n) + \mathbb{E}[\log p_{e,m}(\mathsf{C}_n)]. \tag{18}$$

Choosing $\eta_n = \frac{1}{E(n)}$, we obtain (15), and the assumption that $E(n) \doteq 1$ implies (14). We obtain (16) by writing $\log \alpha = \rho \log(\alpha^{1/\rho})$, writing $\mathbb{E}[\,\cdot\,] = \mathbb{E}[\mathbb{E}[\,\cdot\,|\boldsymbol{X}^{(m)}]]$, and applying Jensen's inequality. ■

The assumption of Theorem 2 is mild, allowing ensembles for which the error probability associated with any two permissible codewords decays nearly *double*-exponentially fast. However, it is a multi-letter condition, and may therefore be difficult to verify directly. A single-letter sufficient condition depending only on the channel, metric and cost constraint (2) is that

$$\lim_{\gamma\to\infty} \frac{1}{\gamma} \log\log \frac{1}{\pi(\gamma)} = 0, \tag{19}$$

where

$$\pi(\gamma) \triangleq \min_{(x,\overline{x})\,:\,c(x)\le\gamma, c(\overline{x})\le\gamma} \mathbb{P}[Y_x \in \mathcal{E}(x,\overline{x})] \tag{20}$$

$$\mathcal{E}(x,\overline{x}) \triangleq \big\{y \,:\, q(\overline{x},y) \ge q(x,y)\big\}, \tag{21}$$

where in (20) we define $Y_x \sim W(\cdot|x)$. Under this assumption, the probability in (13) is lower bounded by the probability that $Y_i \in \mathcal{E}(X_i, \overline{X}_i)$ for $i = 1,\dots,n$, which in turn is lower bounded by $\pi(n\Gamma)^n$. Since n times a subexponential sequence is also subexponential, the condition of Theorem 2 follows from (19). Further discussion is given in Appendix A, along with some examples.

From (15), we can see the advantage of the expurgated ensemble over the non-expurgated one. The former yields the exponent corresponding to $-\frac{1}{n}\mathbb{E}[\log p_{e,m}(\mathsf{C}_n)]$, which is higher in general than that of $-\frac{1}{n}\log\mathbb{E}[p_{e,m}(\mathsf{C}_n)]$ due to Jensen's inequality.

Using L'Hôpital's rule, it is easily shown that $\lim_{\rho\to\infty}\rho\log\mathbb{E}[Z^{1/\rho}] = \mathbb{E}[\log Z]$ for any random variable Z. It follows that the inequality in (16) is actually an equality in the limit as $\rho\to\infty$. At first glance, it may appear that a similar argument can be used to show that (6) yields the same exponent as (15). However, there is an issue with the order of the limits of n and ρ. If we take $\rho\to\infty$ in (6), the factor 2^ρ makes the right-hand side equal ∞. Letting ρ grow slowly with n is also potentially problematic, since the random variable $p_{e,m}(\mathsf{C})$ varies with n.

The bounds in Theorem 2 will prove useful for deriving improved exponents compared to Theorem 1 for some codeword distributions, and for extending the type class enumeration method beyond the finite-alphabet setting.

III. Expurgated Ensembles and Exponents

In this section, we present an overview of various expurgated exponents and the connections between them. Our focus here is primarily on existing exponents or simple variations thereof, though we also provide a dual form of the exponent in [6] which is new to the best of our knowledge. Further exponents which appear for the first time in this paper are given in Theorems 5 and 7 in Section IV.

Throughout the paper, we consider three expurgated ensembles, each of which depends on an input distribution Q:

1) The i.i.d. ensemble is characterized by
$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i). \tag{22}$$
This codeword distribution is valid for both discrete and continuous alphabets, but it is not suitable for channels with cost constraints, since in all non-trivial cases there is a non-zero probability of violating the constraint.
2) The constant-composition ensemble is characterized by
$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{|T^n(Q_n)|}\mathbb{1}\big\{\boldsymbol{x}\in T^n(Q_n)\big\}, \tag{23}$$
where Q_n is a type with the same support as Q such that $|Q_n(x) - Q(x)| = O\big(\frac{1}{n}\big)$ for all x. This codeword distribution relies on the input being finite. It is directly applicable to channels with cost constraints, since each codeword satisfies (2) provided that $\mathbb{E}_{Q_n}[c(X)] \le \Gamma$, which in turn can be achieved provided that $\mathbb{E}_Q[c(X)] \le \Gamma$.

3) The cost-constrained ensemble is characterized by

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\mu_n} \prod_{i=1}^{n} Q(x_i) \mathbb{1}\{\boldsymbol{x} \in \mathcal{D}_n\}, \tag{24}$$

where

$$\mathcal{D}_n \triangleq \left\{ \boldsymbol{x} \,:\, \frac{1}{n}\sum_{i=1}^{n} c(x_i) \le \Gamma, \left| \frac{1}{n}\sum_{i=1}^{n} a_l(x_i) - \phi_l \right| \le \frac{\delta}{n},\, l = 1, \ldots, L \right\}, \tag{25}$$

and where δ is a positive constant (independent of n), $\{a_l(\cdot)\}_{l=1}^{L}$ are functions with means $\phi_l \triangleq \mathbb{E}_Q[a_l(X)]$, and μ_n is a normalizing constant. This codeword distribution is valid for both discrete and continuous alphabets, and ensures that each codeword satisfies (2). Both $c(\cdot)$ and $\{a_l(\cdot)\}$ can be thought of as cost functions, and we will distinguish between the two by referring to them as the *system cost* and *auxiliary costs* respectively. In contrast to the system cost, the auxiliary costs are functions which can be optimized. That is, while the system cost is given as part of the problem statement, the auxiliary costs are introduced to improve the performance of the random-coding ensemble itself [12], [13], [21].

We proceed by stating and comparing the exponents obtained by the above ensembles; derivations will be given in Section IV. Except where stated otherwise, we assume that the channel is a DMC with unconstrained inputs.

A straightforward generalization of Gallager's i.i.d. exponent to the setting of mismatched decoding is as follows:

$$E_{\mathrm{ex}}^{\mathrm{iid}}(Q,R) \triangleq \sup_{\rho \ge 1} E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho) - \rho R, \tag{26}$$

where

$$E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho) \triangleq \sup_{s \ge 0} -\rho \log \sum_{x,\overline{x}} Q(x)Q(\overline{x}) \left(\sum_{y} W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)} \right)^{s} \right)^{1/\rho}. \tag{27}$$

The objective in (27) is concave in s, and under ML decoding (i.e. $q(x,y) = W(y|x)$), it is also unchanged when s is replaced by $1-s$. From these properties, it follows that $s = \frac{1}{2}$ is optimal for ML decoding, and thus the exponent is the same as that of Gallager [2].

Csiszár and Körner [6] make use of the constant-composition codeword distribution in (23). The analysis is significantly different to that of Gallager, and yields an exponent in a different form, namely

$$E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R) \triangleq \min_{\substack{P_{X\overline{X}Y} \in \mathcal{T}(Q) \\ I_P(X;\overline{X}) \le R}} D(P_{X\overline{X}Y} \| Q \times Q \times W) - R, \tag{28}$$

where the notation $Q \times Q \times W$ denotes the distribution $Q(x)Q(\overline{x})W(y|x)$, and

$$\mathcal{T}(Q) \triangleq \left\{ P_{X\overline{X}Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{Y}) \,:\, P_X = Q, P_{\overline{X}} = Q, \mathbb{E}_P[\log q(\overline{X},Y)] \ge \mathbb{E}_P[\log q(X,Y)] \right\}. \tag{29}$$

The objective in (28) follows from [6, Eq. (32)] and the identity

$$D(P_{X\overline{X}Y} \| Q \times Q \times W) = D(P_{X\overline{X}Y} \| P_{X\overline{X}} \times W) + I_P(X;\overline{X}), \tag{30}$$

which holds for any $P_{X\overline{X}Y}$ such that $P_X = P_{\overline{X}} = Q$. Defining $P_Y(y) \triangleq \sum_x Q(x)W(y|x)$, we observe that $E_{\mathrm{ex}}^{\mathrm{cc}}$ is positive for sufficiently small R provided that $\mathbb{E}_{Q\times W}[\log q(X,Y)] > \mathbb{E}_{Q\times P_Y}[\log q(X,Y)]$. It was shown in [11] that the mismatched capacity is in fact zero unless this condition holds for some Q.

The following theorem provides the means for comparing the above two exponents, as well as that of [4].

Theorem 3. *For any input distribution Q and rate R, we have*

$$E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R) = \sup_{s\geq 0} \min_{\substack{P_{X\overline{X}}\,:\,P_X=Q,P_{\overline{X}}=Q,\\ I_P(X;\overline{X})\leq R}} \mathbb{E}_P[d_s(X,\overline{X})] + I_P(X;\overline{X}) - R \tag{31}$$

$$= \sup_{\rho\geq 1} E_{\mathrm{x}}^{\mathrm{cc}}(Q,\rho) - \rho R, \tag{32}$$

where

$$d_s(x,\overline{x}) \triangleq -\log \sum_y W(y|x)\left(\frac{q(\overline{x},y)}{q(x,y)}\right)^s \tag{33}$$

$$E_{\mathrm{x}}^{\mathrm{cc}}(Q,\rho) \triangleq \sup_{s\geq 0, a(\cdot)} -\rho \sum_x Q(x) \log \sum_{\overline{x}} Q(\overline{x}) \left(\sum_y W(y|x)\left(\frac{q(\overline{x},y)}{q(x,y)}\right)^s \frac{e^{a(\overline{x})}}{e^{a(x)}} \right)^{1/\rho}. \tag{34}$$

Proof: See Appendix B. ∎

Equations (32) and (34) strongly resemble (26)–(27). Equation (31) is a generalization of the exponent in [4], which is recovered by setting $q(x,y) = W(y|x)$ and $s = \frac{1}{2}$. Using the same argument as the one following (27), it can be shown that the latter choice is optimal. From the proof of Theorem 3, this implies the optimality of $s = \frac{1}{2}$ in (34) under ML decoding, though the optimal choice of $a(\cdot)$ is unclear in general. To our knowledge, the expression in (34) has not appeared previously even for ML decoding.

As noted in [6], [16], we can write (31) in the language of rate-distortion theory [22, Ch. 10]. Fix $s \geq 0$ and define

$$D_s(Q,R) \triangleq \min_{\substack{P_{X\overline{X}}\,:\,P_X=Q,P_{\overline{X}}=Q,\\ I_P(X;\overline{X})\leq R}} \mathbb{E}_P[d_s(X,\overline{X})]. \tag{35}$$

This can be interpreted as the distortion-rate function of a source X with a reproduction variable $\overline{X}$, subject to the additional constraint that each reproduction codeword $\overline{\boldsymbol{x}}$ has empirical distribution Q. For any $s \geq 0$, the constraint on the mutual information in (31) is active for sufficiently small R. The supremum of all such rates is given by

$$R_s(Q) \triangleq I_{P^*}(X;\overline{X}), \tag{36}$$

where

$$P^*_{X\overline{X}} \triangleq \operatorname*{arg\,min}_{P_{X\overline{X}}\,:\,P_X=Q,P_{\overline{X}}=Q} \mathbb{E}_P[d_s(X,\overline{X})] + I_P(X;\overline{X}). \tag{37}$$

For $R \leq R_s$ we have $I_P(X;\overline{X}) = R$ under the minimizing $P_{X\overline{X}Y}$, whereas for $R \geq R_s$ the minimum in (31) decreases linearly with R for any fixed s. It follows that

$$E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R) = \sup_{s\geq 0} E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R,s), \tag{38}$$

where

$$E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R,s) \triangleq \begin{cases} D_s(Q,R) & R \leq R_s(Q) \\ D_s(Q,R_s) + R_s(Q) - R & R > R_s(Q). \end{cases} \tag{39}$$

By applying Jensen's inequality to (34) and setting $a(x) = 0$, we immediately obtain

$$E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R) \geq E_{\mathrm{ex}}^{\mathrm{iid}}(Q,R). \tag{40}$$

It was shown in [5, Ex. 10.18] that (40) holds with equality under ML decoding with an optimized input distribution Q. However, when either the decoding rule or input distribution is fixed, the inequality in (40) can be strict; an example is given at the end of this section. In Section IV-A, we show that the stronger exponent $E_{\rm ex}^{\rm cc}$, in the form given in (32), remains achievable in the case of continuous alphabets, with the summations in (34) replaced by integrals. This is proved using the cost-constrained ensemble in (24).

The following proposition generalizes Gallager's expression for the expurgated exponent as $R \to 0^+$ for channels whose zero-error capacity [23] is zero, and shows that the inequality in (40) becomes an equality in the limit.

Proposition 1. *Fix any input distribution Q such that all pairs $(x, \overline{x})$ with $Q(x)Q(\overline{x}) > 0$ share a common output, i.e. $W(y|x)W(y|\overline{x}) > 0$ for some y. Then*

$$\lim_{R \to 0^+} E_{\rm ex}^{\rm cc}(Q, R) = \lim_{R \to 0^+} E_{\rm ex}^{\rm iid}(Q, R) = \sup_{s \geq 0} \mathbb{E}[d_s(X, \overline{X})], \tag{41}$$

where d_s is defined in (33)*, and the expectation is taken with respect to $Q(x)Q(\overline{x})$.*

Proof: See Appendix C. ∎

We conclude this section with a numerical example. The channel is defined by the entries of the $|\mathcal{X}| \times |\mathcal{Y}|$ matrix

$$\begin{bmatrix} 1-2\delta_0 & \delta_0 & \delta_0 \\ \delta_1 & 1-2\delta_1 & \delta_1 \\ \delta_2 & \delta_2 & 1-2\delta_2 \end{bmatrix}, \tag{42}$$

and the decoding metric is defined similarly with a fixed $\delta \in (0, \frac{1}{3})$ in place of each δ_i ($i = 1, 2, 3$), yielding a minimum Hamming distance rule. Figure 1 plots the exponents in the case that $\delta_0 = 0.01$, $\delta_1 = 0.05$, $\delta_2 = 0.25$ and $Q = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. We observe that $E_{\rm ex}^{\rm cc} > E_{\rm ex}^{\rm iid}$ at all positive rates, and the gap is particularly significant in the mismatched case. However, consistent with Proposition 1, the two coincide in the limit as $R \to 0$.

As noted in [6], if Q is optimized, then the two exponents coincide for ML decoding. However, the strict inequality $E_{\rm ex}^{\rm cc} > E_{\rm ex}^{\rm iid}$ remains possible for other decoding rules.

IV. Derivations of the Expurgated Exponents

In this section, we provide several techniques for deriving the expurgated exponents, including those introduced in Section III and a further two in Theorems 5 and 7 below. The approaches given here have various advantages which were outlined in Section I-C, and which are discussed further in Section IV-D. Throughout the section, expectations are written using summations for notational simplicity (e.g. $\mathbb{E}_Q[f(X)] = \sum_x Q(x)f(x)$). However, we will highlight that certain results apply in the case of continuous alphabets upon replacing the summations by integrals.

A. Derivations Using Theorem 1

1) i.i.d. ensemble: We immediately obtain the exponent in (26), as well as its generalization to continuous alphabets, by substituting the i.i.d. distribution in (22) into $\mathrm{rcux}_{\rho,s}$ in (9).

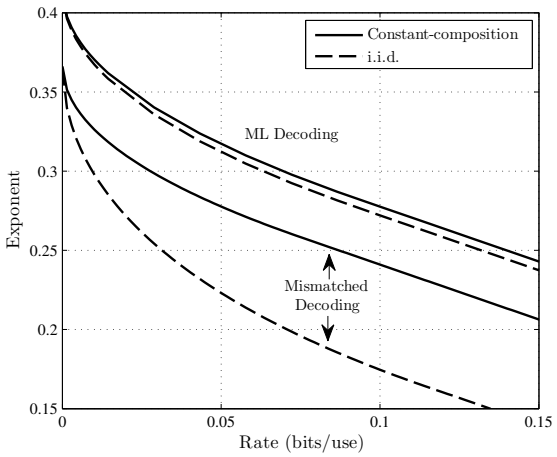

Figure 1. Expurgated exponents for the channel described in (42) with minimum Hamming distance decoding and ML decoding. The parameters are $\delta_0 = 0.01$, $\delta_1 = 0.05$, $\delta_2 = 0.25$ and $Q = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$

2) Constant-composition Ensemble: In the case of finite alphabets, the method of types [5, Ch. 2] can be used to obtain the exact exponents corresponding to rcux_ρ and $\mathrm{rcux}_{\rho,s}$ for each of the ensembles defined in (22)–(24). The analysis is similar for each of these, so we focus on the constant-composition ensemble described by (23). We define

$$\mathcal{S}(Q) \triangleq \left\{\widetilde{P}_{X\overline{X}} \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) \,:\, \widetilde{P}_X = Q, \widetilde{P}_{\overline{X}} = Q\right\} \tag{43}$$

$$\mathcal{T}(\widetilde{P}_{X\overline{X}}) \triangleq \left\{P_{X\overline{X}Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{Y}) \,:\, P_{X\overline{X}} = \widetilde{P}_{X\overline{X}}, \mathbb{E}_P[\log q(\overline{X}, Y)] \geq \mathbb{E}_P[\log q(X, Y)]\right\} \tag{44}$$

$$\mathcal{S}_n(Q) \triangleq \mathcal{S}(Q) \cap \mathcal{P}_n(\mathcal{X}\times\mathcal{X}) \tag{45}$$

$$\mathcal{T}_n(\widetilde{P}_{X\overline{X}}) \triangleq \mathcal{T}(\widetilde{P}_{X\overline{X}}) \cap \mathcal{P}_n(\mathcal{X}\times\mathcal{X}\times\mathcal{Y}). \tag{46}$$

where we overload the symbol $\mathcal{T}$ (see (29)). It follows that $P_{X\overline{X}Y} \in \mathcal{T}(Q)$ (defined in (29)) if and only if $P_{X\overline{X}Y} \in \mathcal{T}(\widetilde{P}_{X\overline{X}})$ (defined in (44)) for some $\widetilde{P}_{X\overline{X}} \in \mathcal{S}(Q)$. We note the following properties of types [5, Ch. 2]:

1) For any $\widetilde{P}_{X\overline{X}} \in \mathcal{S}_n(Q_n)$,

$$\mathbb{P}\big[(\boldsymbol{X}, \overline{\boldsymbol{X}}) \in T^n(\widetilde{P}_{X\overline{X}})\big] \doteq e^{-nI_{\widetilde{P}}(X;\overline{X})}. \tag{47}$$

2) If $(\boldsymbol{x}, \overline{\boldsymbol{x}}) \in T^n(\widetilde{P}_{X\overline{X}})$, then for any $P_{X\overline{X}Y} \in \mathcal{T}_n(\widetilde{P}_{X\overline{X}})$,

$$\mathbb{P}\big[(\boldsymbol{x}, \overline{\boldsymbol{x}}, \boldsymbol{Y}) \in T^n(P_{X\overline{X}Y}) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \doteq e^{-nD(P_{X\overline{X}Y}\|\widetilde{P}_{X\overline{X}}\times W)}. \tag{48}$$

Theorem 4. *Consider a discrete memoryless channel, and let the codeword distribution $P_{\boldsymbol{X}}$ be the constant-composition distribution in* (23) *for some input distribution Q. The bound* rcux_ρ *in* (8) *satisfies the following for any rate $R > 0$:*

$$\inf_{\rho\geq 1} \mathrm{rcux}_\rho(n, e^{nR}) \doteq e^{-nE_{\mathrm{ex}}^{\mathrm{cc}}(Q,R)}. \tag{49}$$

Proof: Using the codeword distribution in (23) and expanding (8) in terms of types, we obtain

$$\begin{aligned}
&\mathrm{rcux}_\rho(n,M)^{1/\rho} \\
&= 4(M-1)\sum_{\widetilde{P}_{X\overline{X}}\in\mathcal{S}_n(Q_n)} \mathbb{P}\Big[(\boldsymbol{X},\overline{\boldsymbol{X}})\in T^n(P_{X\overline{X}})\Big] \sum_{P_{X\overline{X}Y}\in\mathcal{T}_n(\widetilde{P}_{X\overline{X}})} \mathbb{P}\Big[(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{Y})\in T^n(P_{X\overline{X}Y})\,\Big|\,\boldsymbol{X}=\boldsymbol{x}\Big]^{1/\rho} && (50)\\
&\doteq M \max_{\widetilde{P}_{X\overline{X}}\in\mathcal{S}_n(Q_n)} \max_{P_{X\overline{X}Y}\in\mathcal{T}_n(\widetilde{P}_{X\overline{X}})} \exp\Big(-nI_{\widetilde{P}}(X;\overline{X})\Big)\exp\Big(-n\cdot\frac{1}{\rho}D\big(P_{X\overline{X}Y}\|\widetilde{P}_{X\overline{X}}\times W\big)\Big) && (51)\\
&\doteq M \max_{P_{X\overline{X}Y}\in\mathcal{T}(Q)} \exp\Big(-n\Big(\frac{1}{\rho}D\big(P_{X\overline{X}Y}\|P_{X\overline{X}}\times W\big)+I_P(X;\overline{X})\Big)\Big), && (52)
\end{aligned}$$

where in (50) we define $(\boldsymbol{x},\overline{\boldsymbol{x}})$ to be an arbitrary pair with joint type $\widetilde{P}_{X\overline{X}}$, (51) follows from the properties of types in (47)–(48) and the fact that the number of joint types is polynomial in n, and (52) follows from the definitions of $\mathcal{S}_n$, $\mathcal{T}_n$ and $\mathcal{T}$, and by using a standard continuity argument to expand the maximization from types to general distributions (e.g. see [24]). We thus obtain the exponent

$$\begin{aligned}
&\sup_{\rho\ge1}\min_{P_{X\overline{X}Y}\in\mathcal{T}(Q)} D(P_{X\overline{X}Y}\|P_{X\overline{X}}\times W)+\rho\big(I_P(X;\overline{X})-R\big) && (53)\\
&\qquad=\min_{P_{X\overline{X}Y}\in\mathcal{T}(Q)}\sup_{\rho\ge1} D(P_{X\overline{X}Y}\|P_{X\overline{X}}\times W)+\rho\big(I_P(X;\overline{X})-R\big), && (54)
\end{aligned}$$

where (54) follows from Fan's minimax theorem [25], the conditions of which are satisfied here since the objective is linear in ρ and convex in $P_{X\overline{X}Y}$. Using

$$\sup_{\rho\ge1}\rho\alpha=\begin{cases}\infty & \alpha>0\\ \alpha & \alpha\le0\end{cases} \tag{55}$$

and the identity in (30), it follows that (54) coincides with (28). ■

The preceding derivation of $E_{\mathrm{ex}}^{\mathrm{cc}}$ provides a simple alternative to that of Csiszár and Körner [6], while yielding the exponent in the same form.

3) Cost-constrained Ensemble: Here we provide a derivation of $E_{\mathrm{ex}}^{\mathrm{cc}}$ in the form given in (34), as well as its generalization to continuous alphabets, using the cost-constrained ensemble in (24). We allow for a system cost constraint of the form given in (2). A key property of the ensemble which will prove useful in the derivations is

$$\boldsymbol{x}\in\mathcal{D}_n \implies e^{r\left(\sum_{i=1}^n a(x_i)-n\phi_a\right)}e^{|r|\delta}\ge1, \tag{56}$$

which holds for any real number r, and follows immediately from (25). Furthermore, we have the following.

Proposition 2. [13, Prop. 1] *Fix any input distribution Q and set of cost functions $\{a_l\}_{l=1}^L$ such that $E_Q[c(X)]\le\Gamma$, $E_Q[c(X)^2]<\infty$ and $E_Q[a_l(X)^2]<\infty$ for $l=1,\dots,L$. Then the normalizing constant μ_n in* (24) *satisfies*

$$\lim_{n\to\infty}\frac{1}{n}\log\mu_n=0. \tag{57}$$

The following theorem gives an achievable error exponent for a fixed set of auxiliary costs.

Theorem 5. *Consider a memoryless (possibly continuous) channel, and fix any input distribution Q and functions $\{a_l\}$ satisfying the assumptions of Proposition 2. Under the cost-constrained distribution in* (24)*, we have*

$$\inf_{\rho\ge1,s\ge0}\mathrm{rcux}_{\rho,s}(n,e^{nR})\doteq e^{-nE_{\mathrm{ex}}^{\mathrm{cost}}(Q,R,\{a_l\})} \tag{58}$$

for any rate $R > 0$, where

$$E_{\mathrm{ex}}^{\mathrm{cost}}(Q,R,\{a_l\}) \triangleq \sup_{\rho\geq 1} E_{\mathrm{x}}^{\mathrm{cost}}(Q,\rho,\{a_l\}) - \rho R, \tag{59}$$

and[1]

$$E_{\mathrm{x}}^{\mathrm{cost}}(Q,R,\{a_l\}) \triangleq \sup_{s\geq 0,\{r_l\},\{\overline{r}_l\}} -\rho \log \sum_{x,\overline{x}} Q(x)Q(\overline{x}) \left(\sum_y W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)} \right)^s \frac{e^{\sum_{l=1}^L \overline{r}_l (a_l(\overline{x})-\phi_l)}}{e^{\sum_{l=1}^L r_l (a_l(x)-\phi_l)}} \right)^{1/\rho}. \tag{60}$$

Proof: Let $a_l^n(\boldsymbol{x}) \triangleq \sum_{i=1}^n a_l(x_i)$ and $Q^n(\boldsymbol{x}) \triangleq \prod_{i=1}^n Q(x_i)$. We start with (9), and write

$$\mathrm{rcux}_{\rho,s}(n,M)^{1/\rho} = 4(M-1) \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}(\boldsymbol{x}) P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}) \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x}) \left(\frac{q^n(\overline{\boldsymbol{x}},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})} \right)^s \right)^{1/\rho} \tag{61}$$

$$\dot{\leq} M \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}(\boldsymbol{x}) P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}) \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x}) \left(\frac{q^n(\overline{\boldsymbol{x}},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})} \right)^s \frac{e^{\sum_{l=1}^L \overline{r}_l (a_l^n(\overline{\boldsymbol{x}})-n\phi_l)}}{e^{\sum_{l=1}^L r_l (a_l^n(\boldsymbol{x})-n\phi_l)}} \right)^{1/\rho} \tag{62}$$

$$\dot{\leq} M \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} Q^n(\boldsymbol{x}) Q^n(\overline{\boldsymbol{x}}) \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x}) \left(\frac{q^n(\overline{\boldsymbol{x}},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})} \right)^s \frac{e^{\sum_{l=1}^L \overline{r}_l (a_l^n(\overline{\boldsymbol{x}})-n\phi_l)}}{e^{\sum_{l=1}^L r_l (a_l^n(\boldsymbol{x})-n\phi_l)}} \right)^{1/\rho}, \tag{63}$$

where (62) holds for any $\{r_l\}$ and $\{\overline{r}_l\}$ from (56), and (63) follows from (24) and Proposition 2. The proof is concluded by expanding each term in (63) as a product from 1 to n and optimizing ρ, s, $\{r_l\}$ and $\{\overline{r}_l\}$. ■

We now show that we can recover $E_{\mathrm{ex}}^{\mathrm{cc}}$ from $E_{\mathrm{ex}}^{\mathrm{cost}}$ upon setting $L = 2$ and optimizing the auxiliary costs; an analogous statement was shown to be true for the random-coding exponent in [13]. Setting $\overline{r}_1 = r_2 = 1$ and $\overline{r}_2 = r_1 = 0$, and optimizing $a_1(\cdot)$ and $a_2(\cdot)$, we obtain

$$E_{\mathrm{x}}^{\mathrm{cost}}(Q,\rho) = \sup_{s\geq 0,a_1(\cdot),a_2(\cdot)} -\rho \log \sum_{x,\overline{x}} Q(x)Q(\overline{x}) \left(\sum_y W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)} \right)^s \frac{e^{a_1(\overline{x})-\phi_1}}{e^{a_2(x)-\phi_2}} \right)^{1/\rho} \tag{64}$$

$$\leq \sup_{s\geq 0,a_1(\cdot),a_2(\cdot)} -\rho \sum_x Q(x) \log \sum_{\overline{x}} Q(\overline{x}) \left(\sum_y W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)} \right)^s \frac{e^{a_1(\overline{x})-\phi_1}}{e^{a_2(x)-\phi_2}} \right)^{1/\rho}, \tag{65}$$

where (65) follows from Jensen's inequality. For any s and $a_1(\cdot)$, there exists a choice of $a_2(\cdot)$ that makes Jensen's inequality hold with equality in (65), and hence the same is true after taking the supremum. Hence, and by writing

$$-\sum_x Q(x) \log \left(\frac{e^{-\phi_1}}{e^{a_2(x)-\phi_2}} \right)^{1/\rho} = -\sum_x Q(x) \log \left(e^{-a_1(x)} \right)^{1/\rho} = \frac{\phi_1}{\rho}, \tag{66}$$

we see that the $a_2(\cdot)$ achieving the supremum in (64) is the one yielding equality in (65). Renaming $a_1(\cdot)$ as $a(\cdot)$ and using the first equality in (66), we obtain (34).

It should be noted that, in accordance with Proposition 2, the supremum over s and $a(\cdot)$ in (34) is restricted to choices such that $E_Q[a(X)^2] < \infty$, and such that $E_Q[a_2(X)^2] < \infty$ for the choice of $a_2(\cdot)$ which makes Jensen's inequality hold with equality in (65) (expressed in terms of s and $a(\cdot)$). This may rule out some parameters in the case of infinite or continuous alphabets.

While the parameters $\{r_l\}$ and $\{\overline{r}_l\}$ are not necessary for obtaining (64), they can improve the exponent for a given set of auxiliary costs [13]. That is, the more general exponent of Theorem 5 serves as an indicator of the performance when the auxiliary costs are chosen suboptimally. Using a similar argument to that of (64)–(66), it is easily shown that $E_{\mathrm{ex}}^{\mathrm{cost}} \leq E_{\mathrm{ex}}^{\mathrm{cc}}$, and hence one cannot improve on the exponent obtained using $L = 2$ optimally chosen auxiliary costs.

[1]In the case of continuous alphabets, the summations over sequences should be replaced by integrals.

B. Derivation Using Type Class Enumerators

In the proof of Theorem 4, we gave an exponentially tight analysis of rcux_ρ. In this subsection, we show that an exponentially tight analysis can be provided starting from an earlier step using the method of type class enumeration (e.g. see [7]–[9]). Once again, the analysis is similar for each of the ensembles in (22)–(24), so we focus on the constant-composition ensemble described by (23).

Substituting (10) into (6) and defining

$$d_q(\boldsymbol{x},\overline{\boldsymbol{x}}) \triangleq -\log \mathbb{P}\big[q^n(\overline{\boldsymbol{x}},\boldsymbol{Y}) \geq q^n(\boldsymbol{x},\boldsymbol{Y}) \,\big|\, \boldsymbol{X}=\boldsymbol{x}\big], \tag{67}$$

we obtain the bound

$$p_e(\mathcal{C}) \leq \big(2A_n(R,\rho)\big)^\rho, \tag{68}$$

where

$$A_n(R,\rho) \triangleq \mathbb{E}\bigg[\Big(\sum_{\overline{m}\neq m} e^{-d_q(\boldsymbol{X}^{(m)},\boldsymbol{X}^{(\overline{m})})}\Big)^{1/\rho}\bigg]. \tag{69}$$

This bound provides the starting point for our analysis. Note that since we have not used the inequality in (12), we may allow for $\rho \geq 0$ rather than just $\rho \geq 1$.

Theorem 6. *Consider a discrete memoryless channel, and let the codeword distribution $P_{\boldsymbol{X}}$ be the constant-composition distribution in* (23) *for some input distribution Q. Then the following holds for any rate $R>0$:*

$$\inf_{\rho\geq 0} \big(2A_n(R,\rho)\big)^\rho \doteq e^{-nE_{\mathrm{ex}}^{\mathrm{cc}}(Q,R)}. \tag{70}$$

Proof: For $m=1,\ldots,M$ and each joint type $\widetilde{P}_{X\overline{X}}$, we define the random variable

$$N_m(\widetilde{P}_{X\overline{X}}) \triangleq \sum_{\overline{m}\neq m} \mathbb{1}\big\{(\boldsymbol{X}^{(m)},\boldsymbol{X}^{(\overline{m})}) \in T^n(\widetilde{P}_{X\overline{X}})\big\}. \tag{71}$$

Under the random-coding distribution in (23), we have $N_m(\widetilde{P}_{X\overline{X}})=0$ with probability one if $\widetilde{P}_{X\overline{X}} \notin \mathcal{S}_n(Q_n)$. That is, the marginal distribution of each codeword must agree with Q. Since d_q depends only on the joint type of its arguments, we define $d_q(\widetilde{P}_{X\overline{X}}) \triangleq \frac{1}{n} d_q(\boldsymbol{x},\overline{\boldsymbol{x}})$, where $(\boldsymbol{x},\overline{\boldsymbol{x}}) \in T^n(\widetilde{P}_{X\overline{X}})$.

Making repeated use of the fact that the number of joint types is polynomial in n, we have the following:

$$A_n(R,\rho) = \mathbb{E}\bigg[\Big(\sum_{\widetilde{P}_{X\overline{X}}} N_m(\widetilde{P}_{X\overline{X}}) e^{-nd_q(\widetilde{P}_{X\overline{X}})}\Big)^{1/\rho}\bigg] \tag{72}$$

$$\doteq \mathbb{E}\Big[\max_{\widetilde{P}_{X\overline{X}}} N_m(\widetilde{P}_{X\overline{X}})^{1/\rho} e^{-nd_q(\widetilde{P}_{X\overline{X}})/\rho}\Big] \tag{73}$$

$$\doteq \mathbb{E}\Big[\sum_{\widetilde{P}_{X\overline{X}}} N_m(\widetilde{P}_{X\overline{X}})^{1/\rho} e^{-nd_q(\widetilde{P}_{X\overline{X}})/\rho}\Big] \tag{74}$$

$$\doteq \max_{\widetilde{P}_{X\overline{X}}} \mathbb{E}\Big[N_m(\widetilde{P}_{X\overline{X}})^{1/\rho}\Big] e^{-nd_q(\widetilde{P}_{X\overline{X}})/\rho}, \tag{75}$$

where (75) follows by first taking the summation outside the expectation. It follows from (75) that

$$\big(2A_n(R,\rho)\big)^\rho \doteq \max_{\widetilde{P}_{X\overline{X}}} \Big(\mathbb{E}\Big[N_m(\widetilde{P}_{X\overline{X}})^{1/\rho}\Big]\Big)^\rho e^{-nd_q(\widetilde{P}_{X\overline{X}})}. \tag{76}$$

Similarly to [7, Eq. (34)], we have for all $\widetilde{P}_{X\overline{X}} \in \mathcal{S}_n(Q_n)$ that

$$\mathbb{E}\Big[N_m(\widetilde{P}_{X\overline{X}})^{1/\rho}\Big] \doteq \begin{cases} \exp\big(n\big(R - I_{\widetilde{P}}(X;\overline{X})\big)\big) & R < I_{\widetilde{P}}(X;\overline{X}) \\ \exp\big(n\big(R - I_{\widetilde{P}}(X;\overline{X})\big)/\rho\big) & R \geq I_{\widetilde{P}}(X;\overline{X}). \end{cases} \tag{77}$$

This follows from the fact that given $\boldsymbol{X}^{(m)} = \boldsymbol{x}$, $N_m(\widetilde{P}_{X\overline{X}})$ is the sum of $e^{nR}-1$ binary independent random variables,

$$U_{\overline{m}} \triangleq \mathbb{1}\Big\{(\boldsymbol{x}, \boldsymbol{X}^{(\overline{m})}) \in T^n(\widetilde{P}_{X\overline{X}})\Big\}, \tag{78}$$

whose expectations are of the exponential order of $e^{-nI_{\widetilde{P}}(X;\overline{X})}$ (see (47)). Furthermore, expanding (67) in terms of types and using the property in (48), we obtain

$$e^{-nd_q(\widetilde{P}_{X\overline{X}})} \doteq \exp\bigg(-n \min_{P_{X\overline{X}Y} \in \mathcal{T}(\widetilde{P}_{X\overline{X}})} D\big(P_{X\overline{X}Y} \| \widetilde{P}_{X\overline{X}} \times W\big)\bigg) \tag{79}$$

$$\triangleq e^{-nD_q(\widetilde{P}_{X\overline{X}})}. \tag{80}$$

Upon taking into account all the possible empirical distributions $\{\widetilde{P}_{X\overline{X}}\}$ in (78), we obtain

$$\big(2A_n(R,\rho)\big)^{\rho} \doteq e^{-n\min\{E_1(R,\rho), E_2(R)\}}, \tag{81}$$

where

$$E_1(R,\rho) \triangleq \min_{\substack{\widetilde{P}_{X\overline{X}} \in \mathcal{S}(Q) \\ I_{\widetilde{P}}(X;\overline{X}) \geq R}} D_q(\widetilde{P}_{X\overline{X}}) + \rho\big(I_{\widetilde{P}}(X;\overline{X}) - R\big) \tag{82}$$

and

$$E_2(R) = \min_{\substack{\widetilde{P}_{X\overline{X}} \in \mathcal{S}(Q) \\ I_{\widetilde{P}}(X;\overline{X}) \leq R}} D_q(\widetilde{P}_{X\overline{X}}) + I_{\widetilde{P}}(X;\overline{X}) - R. \tag{83}$$

Combining (30), (80) and (83), we see that $E_2(R)$ coincides with $E_{\mathrm{ex}}^{\mathrm{cc}}$ in the form given in (28). It remains to show that $E_1(R,\rho)$, for the optimum choice of ρ, is never smaller than $E_2(R)$. This can be seen by noting that since (82) contains the constraint $I_{\widetilde{P}}(X;\overline{X}) \geq R$, the term multiplying ρ in (82) is non-negative. Thus, the best choice of ρ is to take the limit as $\rho \to \infty$, and hence the minimum in (82) is achieved by some $\widetilde{P}_{X\overline{X}}$ satisfying $I_{\widetilde{P}}(X;\overline{X}) = R$. Since this joint distribution also satisfies the constraints in (83), we conclude that $E_1 \geq E_2$, thus completing the proof. ■

While the exponents of Theorems 4 and 6 coincide for the constant-composition ensemble, the type enumeration approach can yield strictly higher exponents for other codeword distributions; see Section IV-D for details.

C. Derivation Using Distance Enumerators

In this subsection, we extend the preceding type enumeration analysis to channels with infinite or continuous alphabets, and then discuss the further extension to channels with memory. We make use of Theorem 2, and we assume that the technical assumption therein is satisfied (see Appendix A for discussion). We fix $s \geq 0$ and make use of d_s in (33) (or its counterpart for continuous outputs with an integral in place of the summation), as well as its multi-letter extension

$$d_s^n(\boldsymbol{x}, \overline{\boldsymbol{x}}) \triangleq \sum_{i=1}^{n} d_s(x_i, \overline{x}_i). \tag{84}$$

Theorem 7. *Consider a memoryless (possibly continuous) channel, and fix any codeword distribution $P_{\boldsymbol{X}}$ satisfying the assumption of Theorem 2. The exponent*

$$E_{\mathrm{ex}}(R) \triangleq \mathbb{E}\left[\inf_{D\,:\,R(D,\boldsymbol{X})\le R} D + R(D,\boldsymbol{X}) - R\right] \tag{85}$$

is achievable for any function $R(D,\boldsymbol{x})$ such that $\mathbb{P}\big[d_s^n(\boldsymbol{x},\overline{\boldsymbol{X}}) < nD\big] \dot{\le} e^{-nR(D,\boldsymbol{x})}$ uniformly in $\boldsymbol{x}$, and such that $R(\cdot,\boldsymbol{x})$ is continuous for any given $\boldsymbol{x}$.

Proof: We claim that (16) implies the following analog of (68) for a sequence of codebook $\mathcal{C}_n$ of rate approaching R:

$$p_e(\mathcal{C}_n) \dot{\le} \exp\Big(\rho\,\mathbb{E}\big[\log A_n(R,\rho,\boldsymbol{X}^{(m)})\big]\Big), \tag{86}$$

where

$$A_n(R,\rho,\boldsymbol{X}^{(m)}) \triangleq \mathbb{E}\left[\left(\sum_{\overline{m}\neq m} e^{[-d_s^n(\boldsymbol{X}^{(m)},\boldsymbol{X}^{(\overline{m})})]^+}\right)^{1/\rho} \,\middle|\, \boldsymbol{X}^{(m)}\right]. \tag{87}$$

In the absence of the $[\cdot]^+$ function in the exponent, this follows directly from the union bound and Markov's inequality, similarly to the proof of Theorem 1. The introduction of the $[\cdot]^+$ function corresponds to instead taking the better of Markov's inequality and the trivial bound $\mathbb{P}[\cdot] \le 1$.[2]

For a fixed transmitted codeword $\boldsymbol{X}^{(m)} = \boldsymbol{x}$, we analyze $A_n(R,\rho,\boldsymbol{x})$ using *distance enumerators*:

$$\sum_{\overline{m}\neq m} e^{-[d_s^n(\boldsymbol{x},\boldsymbol{X}^{(\overline{m})})]^+} \le \sum_{k=0}^{\infty} e^{-nk\delta} N_m(k,\boldsymbol{x}), \tag{88}$$

where $\delta > 0$ is arbitrary, and

$$N_m(0,\boldsymbol{x}) \triangleq \sum_{\overline{m}\neq m} \mathbb{1}\big\{d_s^n(\boldsymbol{x},\boldsymbol{X}^{(\overline{m})}) < n\delta\big\} \tag{89}$$

$$N_m(k,\boldsymbol{x}) \triangleq \sum_{\overline{m}\neq m} \mathbb{1}\big\{nk\delta \le d_s^n(\boldsymbol{x},\boldsymbol{X}^{(\overline{m})}) < n(k+1)\delta\big\} \quad (k \ge 1). \tag{90}$$

Using Markov's inequality, we can upper-bound the left-hand side of (13) by $e^{-d_s^n(\boldsymbol{x},\overline{\boldsymbol{x}})}$. It thus follows from the assumption of Theorem 2 that the highest value of k,

$$k_{\max}(n) \triangleq \max_{\boldsymbol{x}\,:\,P_{\boldsymbol{X}}(\boldsymbol{x})>0} \max\big\{k \,:\, \mathbb{P}\big[N_m(k,\boldsymbol{x}) > 0\big] \neq 0\big\}, \tag{91}$$

grows subexponentially in n for all $s \ge 0$. Thus, analogously to (76), the quantity $A_n(R,\rho,\boldsymbol{x})$ defined in (87) satisfies

$$A_n(R,\rho,\boldsymbol{x})^\rho \dot{\le} \max_{k\ge 0}\Big(\mathbb{E}\big[N_m(k,\boldsymbol{x})^{1/\rho}\big]\Big)^\rho e^{-nk\delta}. \tag{92}$$

We further upper bound this expression by removing the lower inequality in the indicator function in (90). The key issue is now to assess the exponential rate of decay of the binary random variable

$$U_{\overline{m}}(\boldsymbol{x}) \triangleq \mathbb{1}\big\{d_s^n(\boldsymbol{x},\boldsymbol{X}^{(\overline{m})}) < n(k+1)\delta\big\} \tag{93}$$

[2]This analysis corrects an error in the conference version of this work [26], where the $[\cdot]^+$ function was omitted. This omission does not affect the analysis for ML decoding, since the Bhattacharyya distance is non-negative. However, in general, the function $d_s(\cdot,\cdot)$ may be negative.

for a given transmitted codeword $\boldsymbol{x}$, i.e. to find the exponent of $\mathbb{P}\big[d_s^n(\boldsymbol{x},\overline{\boldsymbol{X}}) < nD\big]$. This can be done using standard large deviations techniques such as the Chernoff bound. Letting $R(D,\boldsymbol{x})$ be as defined in the theorem statement, we have similarly to (81) that

$$A_n(R,\rho,\boldsymbol{x})^\rho \doteq e^{-n\min\{E_1(R,\rho,\delta,\boldsymbol{x}),E_2(R,\delta,\boldsymbol{x})\}}, \tag{94}$$

where

$$E_1(R,\rho,\delta,\boldsymbol{x}) \triangleq \min_{k\,:\,R((k+1)\delta,\boldsymbol{x})\geq R} k\delta + \rho\big(R((k+1)\delta,\boldsymbol{x}) - R\big) \tag{95}$$

$$E_2(R,\delta,\boldsymbol{x}) \triangleq \min_{k\,:\,R((k+1)\delta,\boldsymbol{x})\leq R} k\delta + R((k+1)\delta,\boldsymbol{x}) - R. \tag{96}$$

Upon taking the limit $\delta\to 0$ and using the assumption that $R(\cdot,\boldsymbol{x})$ is lower semicontinuous, these become

$$E_1(R,\rho,\boldsymbol{x}) \triangleq \inf_{D\,:\,R(D,\boldsymbol{x})\geq R} D + \rho\big(R(D,\boldsymbol{x}) - R\big) \tag{97}$$

$$E_2(R,\boldsymbol{x}) \triangleq \inf_{D\,:\,R(D,\boldsymbol{x})\leq R} D + R(D,\boldsymbol{x}) - R. \tag{98}$$

Analogously to Section IV-B, the optimal choice of ρ is in the limit as $\rho\to\infty$, and we obtain $E_2\leq E_1$, and hence

$$\inf_{\rho\geq 0} A_n(R,\rho,\boldsymbol{x})^\rho \doteq e^{-nE_2(R,\boldsymbol{x})}. \tag{99}$$

Substituting (99) into (86), we obtain $p_e(\mathcal{C}) \doteq e^{-n\mathbb{E}[E_2(R,\boldsymbol{X})]}$, thus yielding (85). ■

After a suitable modification of the definition of $d_s^n(\boldsymbol{x},\overline{\boldsymbol{x}})$, (85) extends immediately to more general channels and metrics (e.g. channels with memory). The ability to simplify the exponent (e.g. to a single-letter expression) depends on the form of $R(D,\boldsymbol{x})$, which in turn depends strongly on the codeword distribution $P_{\boldsymbol{X}}$. In some cases, $P_{\boldsymbol{X}}$ can be chosen in such a way that $R(D,\boldsymbol{x})$ is the same for all $\boldsymbol{x}$ with $P_{\boldsymbol{X}}(\boldsymbol{x})>0$, thus greatly simplifying (85).

In Appendix D, we particularize Theorem 7 to the cost-constrained ensemble with a single auxiliary cost $a_1(x) = a(x)$, and show that after optimizing $a(\cdot)$, (85) yields the exponent $E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R)$ in (32). In accordance with Proposition 2, we require the auxiliary cost $a(\cdot)$ to satisfy $\mathbb{E}_Q[a(X)^2]<\infty$.

D. Comparison of Techniques

For the constant-composition codeword distribution, the approaches of Sections IV-A and IV-B led to the same exponent, namely $E_{\mathrm{ex}}^{\mathrm{cc}}$. It should be noted, however, that the type enumeration approach can yield a strictly higher exponent than that of rcux_ρ in Theorem 1 for some codeword distributions. Here we discuss the simple example of the i.i.d. distribution in (22). Applying properties of types in the same way in Section IV-A, it is easily verified that the exponent of rcux_ρ is

$$\min_{\substack{P_{X\overline{X}Y}\,:\,D(P_{X\overline{X}}\|Q\times Q)\leq R,\\ \mathbb{E}_P[\log q(\overline{X},Y)]\geq\mathbb{E}_P[\log q(X,Y)]}} D(P_{X\overline{X}Y}\|Q\times Q\times W) - R. \tag{100}$$

On the other hand, the analysis of Section IV-B yields an exponent of the same form as (100) with an additional constraint $P_X = Q$ in the minimization. To see this, we note that the quantity $N_m(\widetilde{P}_{X\overline{X}})$ defined in (71) satisfies

$$\mathbb{E}\Big[N_m(\widetilde{P}_{X\overline{X}})^{1/\rho}\Big] = \mathbb{P}\big[\boldsymbol{X}^{(m)}\in T^n(\widetilde{P}_X)\big]\,\mathbb{E}\Big[N_m(\widetilde{P}_{X\overline{X}})^{1/\rho}\,\Big|\,\boldsymbol{X}^{(m)}\in T^n(\widetilde{P}_X)\Big] \tag{101}$$

$$\doteq \begin{cases} \exp\big(-nD(\widetilde{P}_X\|Q)\big)\cdot\exp\big(n\big(R-D(\widetilde{P}_{X\overline{X}}\|\widetilde{P}_X\times Q)\big)\big) & R<I_{\widetilde{P}}(X;\overline{X})\\ \exp\big(-nD(\widetilde{P}_X\|Q)\big)\cdot\exp\big(n\big(R-D(\widetilde{P}_{X\overline{X}}\|\widetilde{P}_X\times Q)\big)/\rho\big) & R\geq I_{\widetilde{P}}(X;\overline{X}). \end{cases} \tag{102}$$

The additional factor $\exp\big(-nD(\widetilde{P}_X\|Q)\big)$ leads to an additive $\rho D(\widetilde{P}_X\|Q)$ term in the exponent E_2 in (83). The optimal choice of ρ is again in the limit as $\rho\to\infty$, and under this choice the minimizing $\widetilde{P}_{X\overline{X}}$ must satisfy $\widetilde{P}_X = Q$ so that the divergence is forced to zero.

Depending on the channel, metric and input distribution, adding the constraint $P_X = Q$ to (100) may yield a strict improvement in the exponent. Since both derivations are exponentially tight from the step at which they start, we conclude that the weakness of the simpler derivation is in the inequality in (11), or more precisely, the use of (12). While this step simplifies the derivations, the above example shows that it is not exponentially tight in general.

Another approach to recovering the constraint $\widetilde{P}_X = Q$ in the above example is to follow the steps of Theorem 1 and Section IV-A starting with Theorem 2. Since the expectation of the transmitted codeword is outside the logarithm in (16), we obtain the constraint $\widetilde{P}_X = Q$ in the final minimization using the fact that the empirical distribution of $\boldsymbol{X}$ is close to Q with high probability. We conclude that the inequality in (12) is exponentially tight for the i.i.d. ensemble when we start with (16), even though it is not tight when we start with (6).

We have provided two derivations of $E_{\mathrm{ex}}^{\mathrm{cc}}$ using the cost-constrained ensemble, namely, those in Sections IV-A and IV-C (along with Appendix D). A notable difference between the derivations is the method for ensuring that the average over x is outside the logarithm in (34), which is desirable due to Jensen's inequality. In Theorem 5, the expectation is inside the logarithm, but the desired result is obtained by choosing $a_2(x)$ to make Jensen's inequality hold with equality. On the other hand, in Appendix D the expectation arises outside the logarithm even in the case that $L = 1$.

Provided that the assumption of Theorem 2 is met, we can combine the two approaches and apply the techniques of Theorem 1 and Section IV-A to (16), in which case $E_{\mathrm{x}}^{\mathrm{cost}}$ in (60) is improved to

$$E_{\mathrm{x}}^{\mathrm{cost}^*}(Q,R,\{a_l\}) \triangleq \sup_{s\geq 0,\{\overline{r}_l\}} -\rho\sum_x Q(x)\log\sum_{\overline{x}} Q(\overline{x})\left(\sum_y W(y|x)\left(\frac{q(\overline{x},y)}{q(x,y)}\right)^s e^{\sum_{l=1}^L \overline{r}_l(a_l(\overline{x})-\phi_l)}\right)^{1/\rho}, \tag{103}$$

where the outer-most summation arises using Proposition 3 in Appendix D. This exponent can also be derived by extending the analysis of Appendix D to include multiple auxiliary costs.

In the case that $L = 0$ (i.e. i.i.d. coding), the Lagrange duality techniques of Theorem 3 reveal that (103) is in fact identical to (100) with the added constraint $P_X = Q$. That is, the additional constraint $P_X = Q$ in the primal expression corresponds to an average over x outside the logarithm in the dual expression.

E. Connections with Statistical Mechanics

It is instructive to look at the analysis of Sections IV-B and IV-C from the statistical-mechanical perspective. Let us take another look at the expression

$$Z(\boldsymbol{x}) = \sum_{\overline{m}\neq m} e^{-d(\boldsymbol{x},\boldsymbol{X}^{(\overline{m})})}, \tag{104}$$

where d can represent either d_q (see (67)) or $[d_s^n]^+$ (see (84)). From the viewpoint of statistical physics, Z can be interpreted as the partition function of a physical system, where for a fixed $\boldsymbol{x}^{(m)} = \boldsymbol{x}$, the various configurations (microstates) are $\{\boldsymbol{x}^{(\overline{m})}\}_{\overline{m}\neq m}$ and the energy function (Hamiltonian) is given by $d(\boldsymbol{x},\overline{\boldsymbol{x}})$. The various "configurational energies" $\{d(\boldsymbol{x},\boldsymbol{X}^{(\overline{m})})\}$ are independent random variables, since the codewords are generated independently. As explained in [27, Ch. 5-6] (see also [9, Ch. 6-7] and references therein), this setting is analogous to the random energy

model (REM) in the literature of statistical physics of magnetic materials. The REM was invented by Derrida [28]–[30] as a model of extremely disordered spin glasses. This model is exactly solvable and exhibits a phase transition: Below a certain critical temperature, the partition function becomes dominated by a subexponential number of configurations in the ground-state energy, which means that the system freezes and its entropy vanishes in the thermodynamic limit. This combination of freezing and disorder resembles the behavior of a glass, so this low temperature phase of zero entropy is called the *glassy phase*. Above the critical temperature, the partition function is dominated by an exponential number of configurations, so its entropy is positive. This high temperature phase is called the *paramagnetic phase.*

In the case that $P_{\boldsymbol{X}}$ is the constant-composition distribution in (23) and $d(\cdot,\cdot)$ represents $[d_s^n(\cdot,\cdot)]^+$, we can link these phases to the exponent $E_{\mathrm{ex}}^{\mathrm{cc}}$ in the form given in (39). The graph of $E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R,s)$ is curved at rates below R_s (see (36)), and is a straight line at rates above R_s. The curved part corresponds to the glassy phase of the REM associated with (104), because the dominant contribution to $\mathbb{E}[Z(\boldsymbol{x})^{1/\rho}]$ (see (104)) is due to a subexponential number of codewords whose "distance" from $\boldsymbol{x}$ (i.e. their "energy") is roughly $nD_s(Q,R)$. The straight-line part, on the other hand, corresponds to the paramagnetic phase, where roughly $e^{n(R-R_s)}$ incorrect codewords at distance $nD_s(Q,R_s)$ dominate the behavior. Thus, the passage between the curved part and the straight-line part at $R=R_s$ can be interpreted as a glassy phase transition. A similar discussion applies for the multi-letter distance d_q used in Section IV-B, with $D_s(Q,R)$ replaced by

$$D_q(Q,R) \triangleq \min_{\widetilde{P}_{X\overline{X}}\in\mathcal{S}(Q)\,:\,I_{\widetilde{P}}(X;\overline{X})\le R} D_q(\widetilde{P}_{X\overline{X}}), \tag{105}$$

where $D_q(\widetilde{P}_{X\overline{X}})$ is defined in (80).

V. Prefactor to the i.i.d. Expurgated Exponent

Error exponents characterize the rate of decay of the error probability as the block length increases. At finite block lengths, the effect of the subexponential prefactor can be significant, and it is therefore of interest to characterize its behavior. There exist several works studying this prefactor for the random-coding exponent [13], [19], [31], [32] and the sphere-packing exponent [31]–[33]. In this section, we characterize the prefactor for the i.i.d. expurgated exponent. We will see that, under some technical conditions, the prefactor to rcux_ρ in (8) behaves as $O\big(\frac{1}{\sqrt{n}}\big)$, thus improving on Gallager's $O(1)$ prefactor. Our analysis builds on that of [13], [20].

A. Preliminary Definitions

We define the sets

$$\mathcal{Y}_1(x,\overline{x}) \triangleq \Big\{y\,:\,W(y|x)W(y|\overline{x})>0\Big\} \tag{106}$$

$$\mathcal{A}(Q) \triangleq \left\{(x,\overline{x})\,:\,Q(x)Q(\overline{x})>0,\ \frac{q(\overline{x},y)}{q(x,y)}\neq\frac{q(\overline{x},y')}{q(x,y')}\ \text{for some } y,y'\in\mathcal{Y}_1(x,\overline{x})\right\} \tag{107}$$

and make the following technical assumptions:

$$q(x,y)=0 \iff W(y|x)=0 \tag{108}$$

$$\mathcal{A}(Q)\neq\emptyset. \tag{109}$$

In the case that $q(x,y) = W(y|x)$ (i.e. ML decoding), (108) is trivial, and (109) reduces to the *non-singularity* assumption of [19]. A notable example where this condition fails is the binary erasure channel (BEC) with $Q = \left(\frac{1}{2}, \frac{1}{2}\right)$.

We write

$$E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho,s) \triangleq -\rho \log \sum_{x,\overline{x}} Q(x)Q(\overline{x}) \left(\sum_{y} W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)} \right)^{s} \right)^{1/\rho} \tag{110}$$

to denote the objective in (27) with a fixed value of s. We define the tiled distribution

$$V_s(y|x,\overline{x}) \triangleq \frac{W(y|x)\left(\frac{q(\overline{x},y)}{q(x,y)}\right)^{s}}{\sum_{y'} W(y'|x)\left(\frac{q(\overline{x},y')}{q(x,y')}\right)^{s}} \tag{111}$$

$$V_s^n(\boldsymbol{y}|\boldsymbol{x},\overline{\boldsymbol{x}}) \triangleq \prod_{i=1}^{n} V_s(y_i|x_i,\overline{x}_i), \tag{112}$$

and the generalized information density

$$j_s(x,\overline{x},y) \triangleq \log \frac{V_s(y|x,\overline{x})}{W(y|x)} \tag{113}$$

$$j_s^n(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{y}) \triangleq \sum_{i=1}^{n} j_s(x_i,\overline{x}_i,y_i). \tag{114}$$

Furthermore, we define the joint tilted distribution

$$P_{\rho,s}^*(x,\overline{x}) = \frac{Q(x)Q(\overline{x})\left(\sum_y W(y|x)\left(\frac{q(\overline{x},y)}{q(x,y)}\right)^{s}\right)^{1/\rho}}{\sum_{x',\overline{x}'} Q(x')Q(\overline{x}')\left(\sum_{y'} W(y'|x')\left(\frac{q(\overline{x}',y')}{q(x',y')}\right)^{s}\right)^{1/\rho}}, \tag{115}$$

and the conditional variance

$$c_0(Q,\rho,s) \triangleq \mathbb{E}\Big[\mathrm{Var}\big[j_s(X_s^*,\overline{X}_s^*,Y_s^*)\big|X_s^*,\overline{X}_s^*\big]\Big], \tag{116}$$

where $(X_s^*,\overline{X}_s^*,Y_s^*) \sim P_{\rho,s}^*(x,\overline{x})V_s(y|x,\overline{x})$. The arguments to c_0 will henceforth be omitted, since their values will be understood from the context.

Writing $Y_s \sim V_s(\cdot|x,\overline{x})$, the following arguments show that the assumptions in (108)–(109) imply that $c_0 > 0$ whenever $s > 0$:

$$\mathrm{Var}[j_s(x,\overline{x},Y_s)] = 0 \iff j_s(x,\overline{x},y) \text{ is independent of } y \text{ wherever } V_s(y|x,\overline{x}) > 0 \tag{117}$$

$$\iff \frac{q(\overline{x},y)}{q(x,y)} \text{ is independent of } y \text{ wherever } W(y|x)q(\overline{x},y) > 0 \tag{118}$$

$$\iff (x,\overline{x}) \notin \mathcal{A}(Q), \tag{119}$$

where (118) follows from the definitions of j_s and V_s, and (119) follows from the assumption in (108) and the definition of $\mathcal{A}(Q)$. Using the assumption in (109), it follows that $c_0 > 0$.

Finally, we define the set

$$\mathcal{I}_s \triangleq \Big\{ j_s(x,x,y) \,:\, W(y|x) > 0, (x,\overline{x}) \in \mathcal{A}(Q) \Big\} \tag{120}$$

and the constant

$$\psi_s \triangleq \begin{cases} 1 & \mathcal{I}_s \text{ does not lie on a lattice} \\ \frac{\overline{h}}{1-e^{-\overline{h}}} & \mathcal{I}_s \text{ lies on a lattice with span } \overline{h}. \end{cases} \tag{121}$$

B. Statement of the Result

Theorem 8. *Fix any DMC W, decoding metric q and input distribution Q satisfying* (108)–(109)*. For any $R > 0$, $\rho \geq 1$ and $s > 0$, there exists a sequence of codebooks $\mathcal{C}_n$ with $M \geq e^{nR}$ codewords whose maximal error probability satisfies*

$$p_e(\mathcal{C}_n) \leq \frac{4^\rho \psi_s}{\sqrt{2\pi n c_0}} e^{-n(E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho,s)-\rho R)}\big(1+o(1)\big) \tag{122}$$

Proof: See Section V-C. ∎

It is interesting to note that under ML coding and any rate where the expurgated exponent and random-coding exponent coincide (i.e. $\rho = 1$ in both cases), Theorem 8 gives the same prefactor growth rate as that of the random-coding exponent [13], [19]. There is an extra factor of four in (122), which can be attributed to the fact that Theorem 8 considers the maximal error rather than the average error. Of course, Theorem 8 is primarily of interest at low rates, where the expurgated exponent exceeds the random-coding exponent.

C. Proof of Theorem 8

The proof makes use of two technical lemmas. The first is a strong large deviations result which was proved in [13], building upon the analysis in the proof of [3, Lemma 47].

Lemma 2. [13, Lemma 1] *Fix $K > 0$, and for each n, let $(n_1, \cdots, n_K)$ be integers such that $\sum_k n_k = n$. Fix the PMFs $Q_1, \cdots, Q_K$ on a finite subset of $\mathbb{R}$, and let $\sigma_1^2, \cdots, \sigma_K^2$ be the corresponding variances. Let $Z_1, \cdots, Z_n$ be independent random variables, n_k of which are distributed according to Q_k for each k. Suppose that $\min_k \sigma_k > 0$ and $\min_k n_k = \Theta(n)$. Defining*

$$\mathcal{I}_0 \triangleq \bigcup_{k\,:\,\sigma_k>0} \big\{z \,:\, Q_k(z) > 0\big\} \tag{123}$$

$$\psi_0 \triangleq \begin{cases} 1 & \mathcal{I}_0 \text{ does not lie on a lattice} \\ \frac{h_0}{1-e^{-h_0}} & \mathcal{I}_0 \text{ lies on a lattice with span } h_0, \end{cases} \tag{124}$$

the summation $S_n \triangleq \sum_i Z_i$ satisfies the following uniformly in t:

$$\mathbb{E}\Big[e^{-S_n}\mathbb{1}\big\{S_n \geq t\big\}\Big] \leq e^{-t}\bigg(\frac{\psi_0}{\sqrt{2\pi V_n}} + o\Big(\frac{1}{\sqrt{n}}\Big)\bigg), \tag{125}$$

where $V_n \triangleq \mathrm{Var}[S_n]$.

The following lemma ensures the existence of a high probability set of $(\boldsymbol{x}, \overline{\boldsymbol{x}})$ pairs such that Lemma 2 can be applied to the inner probability in (8).

Lemma 3. *For any $R > 0$, $\rho \geq 1$, $s > 0$ and (W, q, Q) satisfying* (108)–(109)*, the sequence of sets*

$$\mathcal{F}_{\rho,s}^n(\delta) \triangleq \Big\{(\boldsymbol{x}, \overline{\boldsymbol{x}}) \,:\, \max_{x,\overline{x}} \Big|\hat{P}_{\boldsymbol{x}\overline{\boldsymbol{x}}}(x,\overline{x}) - P_{\rho,s}^*(x,\overline{x})\Big| \leq \delta\Big\} \tag{126}$$

satisfies the following properties:

1) *For any $\delta > 0$ and $(\boldsymbol{x}, \overline{\boldsymbol{x}}) \in \mathcal{F}_{\rho,s}^n(\delta)$, the random variable $\boldsymbol{Y}_s \sim V_s^n(\cdot|\boldsymbol{x}, \overline{\boldsymbol{x}})$ satisfies*

$$\mathrm{Var}[j_s^n(\boldsymbol{x}, \overline{\boldsymbol{x}}, \boldsymbol{Y}_s)] \geq n(c_0 - r(\delta)), \tag{127}$$

where $r(\delta) \to 0$ as $\delta \to 0$.

2) *For any $\delta > 0$, we have*

$$\liminf_{n\to\infty} -\frac{1}{n} \frac{\sum_{(\boldsymbol{x},\overline{\boldsymbol{x}})\notin\mathcal{F}^n_{\rho,s}(\delta)} Q^n(\boldsymbol{x})Q^n(\overline{\boldsymbol{x}})\Big(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\Big(\frac{q(\overline{\boldsymbol{x}},\boldsymbol{y})}{q(\boldsymbol{x},\boldsymbol{y})}\Big)^s\Big)^{1/\rho}}{\sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} Q^n(\boldsymbol{x})Q^n(\overline{\boldsymbol{x}})\Big(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\Big(\frac{q(\overline{\boldsymbol{x}},\boldsymbol{y})}{q(\boldsymbol{x},\boldsymbol{y})}\Big)^s\Big)^{1/\rho}} > 0. \tag{128}$$

Proof of Theorem 8 Based on Lemma 3: Using the bound rcux_ρ in Theorem 1 with the i.i.d. codeword distribution $P_{\boldsymbol{X}} = Q^n$, we have for any $\delta > 0$ that

$$\frac{1}{4(M-1)}\mathrm{rcux}_\rho(n,M)^{1/\rho} = \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} Q^n(\boldsymbol{x})Q^n(\overline{\boldsymbol{x}})\mathbb{P}\Big[q^n(\overline{\boldsymbol{x}},\boldsymbol{Y}) \ge q^n(\boldsymbol{x},\boldsymbol{Y})\Big]^{1/\rho} \tag{129}$$

$$= \sum_{(\boldsymbol{x},\overline{\boldsymbol{x}})\in\mathcal{F}^n_{\rho,s}(\delta)} Q^n(\boldsymbol{x})Q^n(\overline{\boldsymbol{x}})\mathbb{P}\Big[q^n(\overline{\boldsymbol{x}},\boldsymbol{Y}) \ge q^n(\boldsymbol{x},\boldsymbol{Y})\Big]^{1/\rho}$$

$$+ \sum_{(\boldsymbol{x},\overline{\boldsymbol{x}})\notin\mathcal{F}^n_{\rho,s}(\delta)} Q^n(\boldsymbol{x})Q^n(\overline{\boldsymbol{x}})\mathbb{P}\Big[q^n(\overline{\boldsymbol{x}},\boldsymbol{Y}) \ge q^n(\boldsymbol{x},\boldsymbol{Y})\Big]^{1/\rho}, \tag{130}$$

where each probability is implicitly conditioned on $\boldsymbol{X} = \boldsymbol{x}$.

We first analyze the summation over $\mathcal{F}^n_{\rho,s}(\delta)$ in (130). In order to make the inner probability more amenable to an application of Lemma 2, we write it as

$$\mathbb{P}\Big[q^n(\overline{\boldsymbol{x}},\boldsymbol{Y}) \ge q^n(\boldsymbol{x},\boldsymbol{Y})\Big] = \mathbb{P}\left[\left(\frac{q^n(\overline{\boldsymbol{x}},\boldsymbol{Y})}{q^n(\boldsymbol{x},\boldsymbol{Y})}\right)^s \ge 1\right] \tag{131}$$

$$= \mathbb{P}\left[\frac{\Big(\frac{q^n(\overline{\boldsymbol{x}},\boldsymbol{Y})}{q^n(\boldsymbol{x},\boldsymbol{Y})}\Big)^s}{\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\Big(\frac{q^n(\overline{\boldsymbol{x}},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})}\Big)^s} \ge \frac{1}{\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\Big(\frac{q^n(\overline{\boldsymbol{x}},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})}\Big)^s}\right] \tag{132}$$

$$= \mathbb{P}\left[j^n_s(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{Y}) \ge -\log\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\left(\frac{q(\overline{\boldsymbol{x}},\boldsymbol{y})}{q(\boldsymbol{x},\boldsymbol{y})}\right)^s\right], \tag{133}$$

where j^n_s is defined in (114). Next, following [34, Sec. 3.4.5], we note that the following holds when $V^n_s(\boldsymbol{y}|\boldsymbol{x},\overline{\boldsymbol{x}}) \ne 0$:

$$W^n(\boldsymbol{y}|\boldsymbol{x}) = W^n(\boldsymbol{y}|\boldsymbol{x})\frac{V^n_s(\boldsymbol{y}|\boldsymbol{x},\overline{\boldsymbol{x}})}{V^n_s(\boldsymbol{y}|\boldsymbol{x},\overline{\boldsymbol{x}})} \tag{134}$$

$$= V^n_s(\boldsymbol{y}|\boldsymbol{x},\overline{\boldsymbol{x}})e^{-nj_s(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{y})}. \tag{135}$$

Summing (135) over all $\boldsymbol{y}$ such that $j_s(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{y}) \ge t$, we obtain

$$\mathbb{P}\big[j^n_s(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{Y}) \ge t\big] = \mathbb{E}\Big[e^{-j^n_s(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{Y}_s)}\mathbb{1}\big\{j^n_s(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{Y}_s) \ge t\big\}\Big], \tag{136}$$

where $\boldsymbol{Y}_s \sim V^n_s(\cdot|\boldsymbol{x},\overline{\boldsymbol{x}})$. For any $(\boldsymbol{x},\overline{\boldsymbol{x}}) \in \mathcal{F}^n_{\rho,s}(\delta)$, we obtain the following using Lemma 2, the first part of Lemma 3, and the fact that $c_0 > 0$ (see the arguments following (117)):

$$\mathbb{P}\big[j^n_s(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{Y}) \ge t\big] \le \frac{\psi_s(1+o(1))}{\sqrt{2\pi n(c_0 - r(\delta))}}e^{-t} \tag{137}$$

uniformly in t, provided that δ is sufficiently small so that $r(\delta) < c_0$. Substituting (137) into (133), we obtain

$$\mathbb{P}\Big[q^n(\overline{\boldsymbol{x}},\boldsymbol{Y}) \ge q^n(\boldsymbol{x},\boldsymbol{Y})\Big] \le \frac{\psi_s(1+o(1))}{\sqrt{2\pi n(c_0 - r(\delta))}}\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\left(\frac{q(\overline{\boldsymbol{x}},\boldsymbol{y})}{q(\boldsymbol{x},\boldsymbol{y})}\right)^s, \tag{138}$$

and hence

$$\sum_{(\boldsymbol{x},\overline{\boldsymbol{x}})\in\mathcal{F}^n_{\rho,s}(\delta)} Q^n(\boldsymbol{x})Q^n(\overline{\boldsymbol{x}})\mathbb{P}\Big[q^n(\overline{\boldsymbol{x}},\boldsymbol{Y})\geq q^n(\boldsymbol{x},\boldsymbol{Y})\Big]^{1/\rho}$$
$$\leq \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} Q^n(\boldsymbol{x})Q^n(\overline{\boldsymbol{x}})\bigg(\frac{\psi_s(1+o(1))}{\sqrt{2\pi n(c_0-r(\delta))}}\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\bigg(\frac{q(\overline{\boldsymbol{x}},\boldsymbol{y})}{q(\boldsymbol{x},\boldsymbol{y})}\bigg)^s\bigg)^{1/\rho}. \tag{139}$$

We observe that the right-hand side of (139) has the same exponent as the denominator of (128). Using Markov's inequality, the summation over $\mathcal{F}^n_{\rho,s}(\delta)^c$ in (130) can be upper bounded by the numerator of (128), and thus the second part of Lemma 3 implies

$$\frac{1}{4(M-1)}\mathrm{rcux}_{\rho,s}(n,M)^{1/\rho}\leq \big(1+o(1)\big)\sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} Q^n(\boldsymbol{x})Q^n(\overline{\boldsymbol{x}})\bigg(\frac{\psi_s(1+o(1))}{\sqrt{2\pi n(c_0-r(\delta))}}\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\bigg(\frac{q(\overline{\boldsymbol{x}},\boldsymbol{y})}{q(\boldsymbol{x},\boldsymbol{y})}\bigg)^s\bigg)^{1/\rho}, \tag{140}$$

and hence

$$\mathrm{rcux}_{\rho,s}(n,M)\leq \frac{4^\rho\psi_s(1+o(1))}{\sqrt{2\pi n(c_0-r(\delta))}}M^\rho\bigg(\sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} Q^n(\boldsymbol{x})Q^n(\overline{\boldsymbol{x}})\bigg(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})\bigg(\frac{q(\overline{\boldsymbol{x}},\boldsymbol{y})}{q(\boldsymbol{x},\boldsymbol{y})}\bigg)^s\bigg)^{1/\rho}\bigg)^\rho \tag{141}$$
$$= \frac{4^\rho\psi_s(1+o(1))}{\sqrt{2\pi n(c_0-r(\delta))}}e^{-n(E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho,s)-\rho R)}, \tag{142}$$

where (142) follows by expanding each term as a product from 1 to n and using the definition of $E_{\mathrm{x}}^{\mathrm{iid}}$. The proof is concluded by taking $\delta\to 0$ (and hence $r(\delta)\to 0$). ■

Proof of Lemma 3: We obtain (127) by expanding the variance as

$$\mathrm{Var}[j_s^n(\boldsymbol{x},\overline{\boldsymbol{x}},\boldsymbol{Y}_s)] = \sum_{i=1}^n \mathrm{Var}[j_s(x_i,\overline{x}_i,Y_{s,i})] \tag{143}$$
$$= \sum_{x,\overline{x}} n\hat{P}_{\boldsymbol{x}\overline{\boldsymbol{x}}}(x,\overline{x})\mathrm{Var}[j_s(x,\overline{x},Y_s)] \tag{144}$$

and substituting the bound in the definition of $\mathcal{F}^n_{\rho,s}(\delta)$ in (126). To prove the second property, we note that a nearly identical argument to Section IV-A (based on types) reveals that the exponent of the denominator of (128) is equal to

$$\min_{P_{X\overline{X}}} D(P_{X\overline{X}}\|Q\times Q)+\frac{1}{\rho}\mathbb{E}_P[d_s(X,\overline{X})], \tag{145}$$

where d_s is defined in (33). Similarly, the exponent of the numerator of (128) is given by

$$\min_{P_{X\overline{X}}\,:\,\max_{x,\overline{x}}|P_{X\overline{X}}(x,\overline{x})-P^*_{\rho,s}(x,\overline{x})|>\delta} D(P_{X\overline{X}}\|Q\times Q)+\frac{1}{\rho}\mathbb{E}_P[d_s(X,\overline{X})]. \tag{146}$$

A straightforward analysis of the Karush-Kuhn-Tucker (KKT) conditions [18, Sec. 5.5.3] reveals that (145) is uniquely minimized by $P^*_{\rho,s}$, defined in (115). On the other hand, $P^*_{\rho,s}$ does not satisfy the constraint in in (146), and thus (146) is strictly higher than (145). ■

VI. Discussion and Conclusion

We have presented asymptotic and non-asymptotic expurgated bounds for channels with a given decoding rule. Several expurgated exponents have been derived, including that of Csiszár and Körner [6] and its generalization to continuous alphabets. The type class enumeration approach has been shown to provide better exponents for some codeword distributions, better guarantees of exponential tightness, and the opportunity for deriving expurgated exponents for

channels with memory. By refining the analysis of the i.i.d. ensemble, we have obtained a bound with a $O\big(\frac{1}{\sqrt{n}}\big)$ prefactor, thus improving on Gallager's $O(1)$ prefactor.

APPENDIX

A. Technical Condition of Theorem 2

We begin by providing an example of a class of continuous channels and metrics satisfying the single-letter condition given in (19). Consider an additive noise channel $Y = X + Z$, and let $q(x, y)$ be any decreasing function of $|y - x|$. If the cost constraint is of the form $c(x) = |x|^\beta$ for some constant β, then $c(x) \le \gamma$ if and only if $|x| \le \gamma^{1/\beta}$. Thus, any two permissible points are separated by a distance of at most $2\gamma^{1/\beta}$, and the single-letter condition is satisfied if the additive noise satisfies $\mathbb{P}[Z > 2\gamma^{1/\beta}] \ge e^{-E'(\gamma)}$ and $\mathbb{P}[Z < -2\gamma^{1/\beta}] \ge e^{-E'(\gamma)}$ for some $E'(\gamma)$ growing subexponentially in γ. In particular, this holds for noise distributions with exponential tails (e.g. Gaussian). On the other hand, if the cost function is logarithmic, say $c(x) = \log(1 + |x|)$, then (19) fails for additive noise distributions with exponential tails, since in this case the limit on the left-hand side of (19) equals a positive constant.

For any DMC whose zero-error capacity [23] is zero, the condition of Theorem 2 is satisfied under ML decoding, since the error probability can only decay exponentially. On the other hand, the condition could fail for sufficiently "bad" metrics (e.g. one for which there exists a pair $(x, \overline{x})$ such that $q(x, y) > q(\overline{x}, y)$ for all y). Furthermore, the condition fails under ML decoding whenever the zero-error capacity is positive and Q has a support which includes two inputs not sharing a common output.

B. Proof of Theorem 3

Using the definitions of $\mathcal{S}$ and $\mathcal{T}$ in (43)–(44), we write (28) as

$$\hat{E}_{\mathrm{ex}}^{\mathrm{cc}}(Q, R) = \min_{\substack{\widetilde{P}_{X\overline{X}} \in \mathcal{S}(Q) \\ I_{\widetilde{P}}(X;\overline{X}) \le R}} \min_{P_{X\overline{X}Y} \in \mathcal{T}(\widetilde{P}_{X\overline{X}})} D(P_{X\overline{X}Y} \| \widetilde{P}_{X\overline{X}} \times W) + I_{\widetilde{P}}(X; \overline{X}) - R, \tag{147}$$

where the objective follows from (30). We will study (147) one minimization at a time.

Step 1: For a given $\widetilde{P}_{X\overline{X}} \in \mathcal{S}(Q)$, $I_{\widetilde{P}}(X; \overline{X}) - R$ is constant, and we thus consider the optimization problem

$$\min_{P_{X\overline{X}Y} \in \mathcal{T}(\widetilde{P}_{X\overline{X}})} D(P_{X\overline{X}Y} \| \widetilde{P}_{X\overline{X}} \times W). \tag{148}$$

The Lagrangian [18, Sec. 5.1.1] is given by

$$\begin{aligned} \mathsf{L}_1 = &\sum_{x,\overline{x},y} P_{X\overline{X}Y}(x, \overline{x}, y) \log \frac{P_{X\overline{X}Y}(x, \overline{x}, y)}{\widetilde{P}_{X\overline{X}}(x, \overline{x}) W(y|x)} \\ &+ s\bigg(\sum_{x,y} P_{XY}(x, y) \log q(x, y) - \sum_{\overline{x},y} P_{\overline{X}Y}(\overline{x}, y) \log q(\overline{x}, y) \bigg) + \sum_{x,\overline{x}} \mu(x, \overline{x}) \Big(\widetilde{P}_{X\overline{X}}(x, \overline{x}) - P_{X\overline{X}}(x, \overline{x}) \Big), \end{aligned} \tag{149}$$

where $s \ge 0$ and $\mu(\cdot, \cdot)$ are Lagrange multipliers. The optimization problem is convex with affine constraints, and thus the optimal value is equal to L_1 for some choice of $P_{X\overline{X}Y}$ and the Lagrange multipliers satisfying the Karush-Kuhn-Tucker (KKT) conditions [18, Sec. 5.5.3].

The simplification of (149) using the KKT conditions uses standard arguments, so we omit some details. Setting $\frac{\partial \mathsf{L}_1}{\partial P_{X\overline{X}Y}(x,\overline{x},y)} = 0$, using the constraint $P_{X\overline{X}} = \widetilde{P}_{X\overline{X}}$ to solve for $\mu(\cdot,\cdot)$, and substituting the resulting expressions back into (149), we obtain

$$\mathsf{L}_1 = -\sum_{x,\overline{x}} \widetilde{P}_{X\overline{X}}(x,\overline{x}) \log \sum_y W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)}\right)^s. \tag{150}$$

Renaming $\widetilde{P}_{X\overline{X}}$ as $P_{X\overline{X}}$, taking the supremum over $s \geq 0$, and adding $I_P(X;\overline{X}) - R$ (see (147)–(148)), we obtain the right-hand side of (31) with the minimum and supremum in the opposite order. Using Fan's minimax theorem [25], we can safely interchange the two.

Since we have taken the supremum over the parameter $s \geq 0$ without verifying that it satisfies the KKT conditions, we have only proved that (31) holds with the equality replaced by an inequality ($\leq$). To prove the reverse inequality, we use the log-sum inequality [22, Thm. 2.7.1] similarly to [10, Appendix A]. For any $P_{X\overline{X}Y} \in \mathcal{T}(\widetilde{P}_{X\overline{X}})$, we have

$$D(P_{X\overline{X}Y} \| \widetilde{P}_{X\overline{X}} \times W) \geq D(P_{X\overline{X}Y} \| \widetilde{P}_{X\overline{X}} \times W) - s \sum_{x,\overline{x},y} P_{X\overline{X}Y}(x,\overline{x},y) \log \frac{q(\overline{x},y)}{q(x,y)} \tag{151}$$

$$= \sum_{x,\overline{x},y} P_{X\overline{X}Y}(x,\overline{x},y) \log \frac{P_{X\overline{X}Y}(x,\overline{x},y)}{\widetilde{P}_{X\overline{X}}(x,\overline{x}) W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)}\right)^s} \tag{152}$$

$$\geq \sum_{x,\overline{x}} P_{X\overline{X}}(x,\overline{x}) \log \frac{1}{\sum_y W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)}\right)^s}, \tag{153}$$

where (151) holds for any $s \geq 0$ from the constraint $\mathbb{E}_P[\log q(\overline{X},Y)] \geq \mathbb{E}_P[\log q(X,Y)]$ in (29), (152) follows from the definition of divergence, and (153) follows using the log-sum inequality [22, Thm. 2.7.1] and the constraint $P_{X\overline{X}} = \widetilde{P}_{X\overline{X}}$. Equation (153) coincides with (150), thus completing the proof of (31).

Step 2: We now turn to the proof of (32). For any fixed $s \geq 0$, the Lagrangian corresponding to (31) is given by

$$\mathsf{L}_2 = -\sum_{x,\overline{x}} P_{X\overline{X}}(x,\overline{x}) \log \sum_y W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)}\right)^s + (1+\lambda) \sum_{x,\overline{x}} P_{X\overline{X}}(x,\overline{x}) \log \frac{P_{X\overline{X}}(x,\overline{x})}{Q(x)Q(\overline{x})} - (1+\lambda)R$$
$$+ \sum_x \nu_1(x) \Big(Q(x) - P_X(x)\Big) + \sum_{\overline{x}} \nu_2(\overline{x}) \Big(Q(\overline{x}) - P_{\overline{X}}(\overline{x})\Big), \tag{154}$$

where $\lambda \geq 0$, $\nu_1(\cdot)$ and $\nu_2(\cdot)$ are Lagrange multipliers. Setting $\frac{\partial \mathsf{L}_2}{\partial P_{X\overline{X}}(x,\overline{x})} = 0$, using the constraint $P_X = Q$ to solve for $\nu_1(\cdot)$, and substituting the resulting expressions back into (154), we obtain

$$\mathsf{L}_2 = -(1+\lambda) \sum_x Q(x) \log \sum_{\overline{x}} Q(\overline{x}) \Bigg(\sum_y W(y|x) \left(\frac{q(\overline{x},y)}{q(x,y)}\right)^s \Bigg)^{\frac{1}{1+\lambda}} e^{\frac{1}{1+\lambda}(\nu_2(\overline{x}) - \nu_2(x))} - (1+\lambda)R. \tag{155}$$

Taking the supremum over $\nu_2(\cdot)$, $s \geq 0$ and $\lambda \geq 0$, we obtain the right-hand side of (32) after suitable renaming.

Once again, we have only proved that (32) holds with an inequality ($\leq$) in place of the equality, and we obtain a matching lower bound similarly to (151)–(153). For any $P_{X\overline{X}} \in \mathcal{S}(Q)$ with $I_{\widetilde{P}}(X;\overline{X}) \leq R$, we can lower bound the

objective in (31) as follows:

$$
\begin{aligned}
&-\sum_{x,\overline{x}} P_{X\overline{X}}(x,\overline{x}) \log \sum_y W(y|x) \biggl(\frac{q(\overline{x},y)}{q(x,y)} \biggr)^s + I_P(X;\overline{X}) - R \\
&\quad \geq -\sum_{x,\overline{x}} P_{X\overline{X}}(x,\overline{x}) \log \sum_y W(y|x) \biggl(\frac{q(\overline{x},y)}{q(x,y)} \biggr)^s + \rho\bigl(I_P(X;\overline{X}) - R\bigr) \qquad (156) \\
&\quad = -\rho \sum_{x,\overline{x}} P_{X\overline{X}}(x,\overline{x}) \log \frac{Q(x)Q(\overline{x}) \Bigl(\sum_y W(y|x) \bigl(\frac{q(\overline{x},y)}{q(x,y)} \bigr)^s e^{a(\overline{x})-\phi_a} \Bigr)^{1/\rho}}{P_{X\overline{X}}(x,\overline{x})} - \rho R \qquad (157) \\
&\quad \geq -\rho \sum_{x} Q(x) \log \sum_{\overline{x}} Q(\overline{x}) \Biggl(\sum_y W(y|x) \Bigl(\frac{q(\overline{x},y)}{q(x,y)} \Bigr)^s e^{a(\overline{x})-\phi_a} \Biggr)^{1/\rho} - \rho R, \qquad (158)
\end{aligned}
$$

where (156) holds for any $\rho \geq 1$ from the constraint $I_{\widetilde{P}}(X;\overline{X}) \leq R$, (157) holds for any function $a(x)$ with mean $\phi_a = \mathbb{E}_Q[a(X)]$ by expanding the logarithm and applying simple manipulations, and (158) follows from the log-sum inequality [22, Thm. 2.7.1] and the constraint $P_X = Q$. Using the definition of ϕ_a and again expanding the logarithm, it is easily shown that (158) is unchanged when $e^{a(\overline{x})-\phi_a}$ is replaced by $\frac{e^{a(\overline{x})}}{e^{a(x)}}$, thus completing the proof.

C. Proof of Proposition 1

The result for the i.i.d. exponent follows similarly to Gallager [2, Sec 5.7], so we only explain the differences. Let $E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho,s)$ be the function $E_{\mathrm{x}}^{\mathrm{iid}}$ in (27), with a fixed value of s rather than a supremum. We claim that

$$
\lim_{R\to 0^+} \sup_{\rho\geq 1, s\geq 0} E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho,s) - \rho R = \sup_{\rho\geq 1, s\geq 0} E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho,s). \qquad (159)
$$

It is easily seen that the left-hand side of (159) cannot exceed the right-hand side, since ρR is positive for any sequence of R values approaching zero from above. It remains to prove the converse. We have for all R that

$$
\sup_{\rho\geq 1, s\geq 0} E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho,s) - \rho R \geq E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho,s) - \rho R. \qquad (160)
$$

Taking $R \to 0$ and then taking the supremum over $s \geq 0$ and $\rho \geq 1$ yields the desired result. The remainder of the proof follows using Gallager's argument: For any fixed s, the supremum over ρ is in the limit as $\rho \to \infty$, and this limit is easily evaluated using L'Hôpital's rule.

The result for the constant-composition exponent follows in the same way using the fact that $\sup_{s,a_1(\cdot),a_2(\cdot)} E_{\mathrm{x}}^{\mathrm{cost}}(Q,\rho,\{a_1,a_2\}) = E_{\mathrm{x}}^{\mathrm{cc}}(Q,\rho)$ (see Section IV-A; in particular, $E_{\mathrm{x}}^{\mathrm{cost}}$ is defined in (60)). Once again, the supremum over ρ is in the limit as $\rho \to \infty$ when the remaining parameters are fixed.

D. Derivation of $E_{\mathrm{ex}}^{\mathrm{cc}}$ Using Theorem 7

Using similar arguments to those in Section IV-A, we can evaluate the lower tail probability of $d_s^n(\boldsymbol{x},\overline{\boldsymbol{X}})$ as follows:

$$
\begin{aligned}
\sum_{\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}) \mathbb{1}\bigl\{ d_s^n(\boldsymbol{x},\overline{\boldsymbol{x}}) \leq nD \bigr\} &\leq \sum_{\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}) e^{t(nD - d_s^n(\boldsymbol{x},\overline{\boldsymbol{x}}))} \qquad (161) \\
&\dot{\leq} \sum_{\overline{\boldsymbol{x}}} Q^n(\overline{\boldsymbol{x}}) e^{t(nD - d_s^n(\boldsymbol{x},\overline{\boldsymbol{x}}))} e^{\overline{r}(a(\overline{\boldsymbol{x}}) - n\phi_a)} \qquad (162) \\
&= e^{n(tD - \overline{r}\phi_a)} \prod_{i=1}^n \sum_{\overline{x}} Q(x) e^{\overline{r}a(\overline{x}) - t d_s(x_i,\overline{x})}, \qquad (163)
\end{aligned}
$$

where (161) holds or any $t \geq 0$ by upper bounding the indicator function, and (162) holds for any $\overline{r}$ using (56) and (57). From (163), we may set

$$R(D, \boldsymbol{x}) = \sup_{t \geq 0, \overline{r}} \overline{r}\phi_a - tD - \frac{1}{n}\sum_{i=1}^{n} \theta(x_i, \overline{r}, t), \tag{164}$$

where

$$\theta(x, \overline{r}, t) \triangleq \log \mathbb{E}_Q\big[e^{\overline{r}a(\overline{X}) - td_s(x, \overline{X})}\big]. \tag{165}$$

Before proceeding, we present the following proposition.

Proposition 3. *Consider the cost-constrained distribution $P_{\boldsymbol{X}}$ in* (24)*, and assume that the input distribution Q and auxiliary costs $\{a_l\}_{l=1}^{L}$ are such that assumptions of Proposition 2 are satisfied. For any function $f : \mathcal{X} \to \mathbb{R}$, we have*

$$\lim_{n\to\infty} \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\bigg] = \mathbb{E}_Q[f(X)] \tag{166}$$

provided that $\mathbb{E}_Q[f(X)]$ exists.

Proof: See Appendix E. ∎

We can now simplify the exponent in (85) as follows:

$$\mathbb{E}\Big[\inf_{D\,:\,R(D,\boldsymbol{X})\leq R} D + R(D, \boldsymbol{X}) - R\Big] \tag{167}$$

$$= \mathbb{E}\Big[\inf_{D}\sup_{\rho\geq 1} D + \rho\big(R(D, \boldsymbol{X}) - R\big)\Big] \tag{168}$$

$$\geq \sup_{\rho\geq 1}\mathbb{E}\Big[\inf_{D} D + \rho\big(R(D, \boldsymbol{X}) - R\big)\Big] \tag{169}$$

$$= \sup_{\rho\geq 1}\mathbb{E}\Big[\inf_{D}\sup_{t\geq 0,\overline{r}} D(1-\rho t) - \rho\Big(-\overline{r}\phi_a + \frac{1}{n}\sum_{i=1}^{n}\theta(X_i, \overline{r}, t) + R\Big)\Big] \tag{170}$$

$$\geq \sup_{\rho\geq 1}\sup_{\overline{r}} -\rho\Big(-\overline{r}\phi_a + \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\theta(X_i, \overline{r}, 1/\rho)\Big] + R\Big) \tag{171}$$

$$\to \sup_{\rho\geq 1}\sup_{\overline{r}} \rho\Big(\overline{r}\phi_a - \mathbb{E}_Q[\theta(X, \overline{r}, 1/\rho)] - R\Big)\Big], \tag{172}$$

where (168) follows from (55), (170) follows from (164), (171) follows by replacing the supremum over $t \geq 0$ by the choice $t = 1/\rho$, and (172) follows from Proposition 3.

Substituting (165) into (172) setting $\overline{r} = \frac{1}{\rho}$, and taking the supremum over $a(\cdot)$, we obtain (34), as desired.

E. Proof of Proposition 3

We first present the proof in the case that there is $L = 1$ auxiliary cost $a(\cdot)$ (with mean ϕ_a) and no system cost constraint, and then discuss the changes required to handle the general case. Throughout the proof, we define $a^n(\boldsymbol{x}) \triangleq \sum_{i=1}^{n} a(x_i)$ and $f^n(\boldsymbol{x}) \triangleq \sum_{i=1}^{n} f(x_i)$. We use summations to denote averaging with respect to Q, but the proof remains valid in the continuous case upon replacing these by integrals.

Let $\boldsymbol{X}$ be the random cost-constrained codeword, and define $\boldsymbol{X}' \sim Q^n(\boldsymbol{x}')$. From (24), we have

$$\frac{1}{n}\mathbb{E}\big[f^n(\boldsymbol{X})\big] = \frac{1}{n}\frac{1}{\mu_n}\mathbb{E}\Big[f^n(\boldsymbol{X}')\mathbb{1}\big\{|a^n(\boldsymbol{X}') - n\phi_a| \leq \delta\big\}\Big]. \tag{173}$$

By a direct differentiation, this is equal to $\frac{d}{d\lambda}\big(\frac{1}{n}\log Z(\lambda)\big)$ evaluated at $\lambda=0$, where

$$Z(\lambda) \stackrel{\Delta}{=} \mathbb{E}\Big[e^{\lambda f^n(\boldsymbol{X}')}\mathbb{1}\big\{|a^n(\boldsymbol{X}')-n\phi_a|\le\delta\big\}\Big]. \tag{174}$$

Expanding the expectation and using the inverse Laplace transform relation

$$\mathbb{1}\{z\ge 0\} = \frac{1}{2\pi j}\int_{u-j\infty}^{u+j\infty}\frac{e^{tz}}{t}dt \tag{175}$$

for $u>0$, we have the following:

$$Z(\lambda) = \sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}')e^{\lambda f^n(\boldsymbol{x}')}\Big(\mathbb{1}\{a^n(\boldsymbol{x}')\le n\phi_a+\delta\}-\mathbb{1}\{a^n(\boldsymbol{x}')\le n\phi_a-\delta\}\Big) \tag{176}$$

$$= \frac{1}{2\pi j}\sum_{\boldsymbol{x}'} Q^n(\boldsymbol{x}')e^{\lambda f^n(\boldsymbol{x}')}\int_{u-j\infty}^{u+j\infty} e^{t(n\phi_a-a^n(\boldsymbol{x}'))}\frac{e^{t\delta}-e^{-t\delta}}{t}dt \tag{177}$$

$$= \frac{1}{2\pi j}\int_{u-j\infty}^{u+j\infty}\frac{e^{t\delta}-e^{-t\delta}}{t}e^{n\phi_a t}\bigg(\sum_{x'}Q(x')e^{-ta(x')+\lambda f(x')}\bigg)^n dt. \tag{178}$$

Denoting the derivative of $Z(\cdot)$ by $Z'(\cdot)$, we have

$$Z'(0) = \frac{n}{2\pi j}\int_{u-j\infty}^{u+j\infty}\frac{e^{t\delta}-e^{-t\delta}}{t}e^{n\phi_a t}\bigg(\sum_{x'}Q(x')e^{-ta(x')}\bigg)^{n-1}\sum_{x'}Q(x')f(x')e^{-ta(x')}dt \tag{179}$$

$$= \frac{n}{2\pi j}\int_{u-j\infty}^{u+j\infty}\frac{e^{t\delta}-e^{-t\delta}}{t}e^{n\phi_a t}\bigg(\sum_{x'}Q(x')e^{-ta(x')}\bigg)^{n}\frac{\sum_{x'}Q(x')f(x')e^{-ta(x')}}{\sum_{x'}Q(x')e^{-ta(x')}}dt. \tag{180}$$

Finally, using the assumption that $\mathbb{E}_Q[a(X)^2]<\infty$ and applying the saddlepoint method [35, Ch. 4-5] (see also [9, Sec. 4.2-4.3]), we obtain

$$\frac{d}{d\lambda}\Big(\frac{1}{n}\log Z(\lambda)\Big)\Big|_{\lambda=0} = \frac{Z'(0)}{Z(0)} \to \frac{\sum_{x'}Q(x')f(x')e^{-t_0a(x')}}{\sum_{x'}Q(x')e^{-t_0a(x')}}, \tag{181}$$

where t_0 is the zero of the derivative (saddlepoint) of the function $h(t)=\phi_a t+\log\mathbb{E}_Q[e^{-ta(X)}]$. Since $\phi_a=\mathbb{E}_Q[a(X)]$ by definition, it is easily verified that $t_0=0$, and thus the right-hand side of (181) equals $\mathbb{E}_Q[f(X)]$, as desired.

In the case of multiple auxiliary costs, the argument is similar, but with $ta(\cdot)$ replaced by $\sum_l t_l a_l(\cdot)$. The system cost $c(x)$ in (25) can be handled similarly provided that $\mathbb{E}_Q[c(X)]\le\Gamma$, which is an assumption of the proposition.

REFERENCES

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

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

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

[4] I. Csiszár, J. Körner, and K. Marton, "A new look at the error exponent of discrete memoryless channels," in *IEEE Int. Symp. Inf. Theory*, Ithaca, NY, 1977.

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[6] ——, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, Jan. 1981.

[7] N. Merhav, "Error exponents of erasure/list decoding revisited via moments of distance enumerators," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4439–4447, Oct. 2008.

[8] R. Etkin, N. Merhav, and E. Ordentlich, "Error exponents of optimum decoding for the interference channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 40–56, 2010.

[9] N. Merhav, "Statistical physics and information theory," *Foundations and Trends in Comms. and Inf. Theory*, vol. 6, no. 1-2, pp. 1–212, 2009.

[10] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[11] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 35–43, Jan. 1995.

[12] A. Ganti, A. Lapidoth, and E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[13] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, May 2014.

[14] F. Jelinek, "Evaluation of expurgated bound exponents," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 501–505, 1968.

[15] R. Blahut, "Composition bounds for channel block codes," *IEEE Trans. Inf. Theory*, vol. 23, no. 6, pp. 656–674, 1977.

[16] J. K. Omura, "Expurgated bounds, Bhattacharyya distance, and rate distortion functions," *Inf. and Control*, vol. 24, no. 4, pp. 358 – 383, 1974.

[17] I. Csiszár, "On the error exponent of source-channel transmission with a distortion threshold," *IEEE Trans. Inf. Theory*, vol. 28, no. 6, pp. 823–828, Nov. 1982.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[19] Y. Altuğ and A. B. Wagner, "Refinement of the random coding bound," 2014, http://arxiv.org/abs/1312.6875.

[20] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "A derivation of the asymptotic random-coding prefactor," in *Allerton Conf. on Comm., Control and Comp.*, Monticello, IL, 2013.

[21] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Cost-constrained random coding and applications," in *Inf. Theory and Apps. Workshop*, San Diego, CA, Feb. 2013.

[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 2001.

[23] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. 2, no. 3, pp. 8–19, Sept. 1956.

[24] A. G. D'yachkov, "Bounds on the average error probability for a code ensemble with fixed composition," *Prob. Inf. Transm.*, vol. 16, no. 4, pp. 3–8, 1980.

[25] K. Fan, "Minimax theorems," *Proc. Nat. Acad. Sci.*, vol. 39, pp. 42–47, 1953.

[26] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Expurgated random-coding ensembles: Exponents, refinements and connections," in *Int. Zurich Sem. on Comms.*, Feb. 2014.

[27] M. Mézard and A. Montanari, *Information, Physics and Computation*. Oxford University Press, 2009.

[28] B. Derrida, "Random-energy model: Limit of a family of disordered models," *Phys. Rev. Lett.*, vol. 45, no. 2, pp. 79–82, 1980.

[29] ——, "The random energy model," *Physics Reports*, vol. 67, no. 1, pp. 29–35, 1980.

[30] ——, "Random-energy model: An exactly solvable model for disordered systems," *Phys. Rev. Lett.*, vol. 24, no. 5, pp. 2613–2626, 1981.

[31] P. Elias, "Coding for two noisy channels," in *Third London Symp. Inf. Theory*, 1955.

[32] R. L. Dobrushin, "Asymptotic estimates of the probability of error for transmission of messages over a discrete memoryless communication channel with a symmetric transition probability matrix," *Theory Prob. Apps.*, vol. 7, no. 3, pp. 270–300, 1962.

[33] Y. Altuğ and A. B. Wagner, "Refinement of the sphere-packing bound: Asymmetric channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1592–1614, March 2013.

[34] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Princeton University, 2010.

[35] N. G. de Bruijn, *Asymptotic Methods in Analysis*. Dover Publications, 1981.