Expurgated Joint Source-Channel Coding Bounds and Error Exponents

Jonathan Scarlett
LIONS, EPFL
jonathan.scarlett@epfl.ch

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

Albert Guillén i Fàbregas
ICREA & Universitat Pompeu Fabra
University of Cambridge
guillen@ieee.org

***Abstract*—This paper studies expurgated random-coding bounds and error exponents for joint source-channel coding (JSCC). We extend Gallager's expurgation techniques for channel coding to the JSCC setting, and derive a non-asymptotic bound that recovers two exponents derived by Csiszár using the method of types. Our approach has the notable advantage of being directly applicable to channels with continuous alphabets.**

I. Introduction

It was shown by Shannon [1] that separate source-channel coding incurs no loss of optimality in terms of asymptotically reliable transmission for discrete memoryless sources and channels. However, joint source-channel coding (JSCC) has been shown to yield performance gains in terms of the error exponent [2]–[6], second-order asymptotics [7], [8], and non-asymptotic performance [6], [8], [9]. In this paper, we consider expurgated random coding for JSCC, which has thus far received significantly less attention than random coding with independent codewords.

In the channel coding setting, the main approaches for obtaining expurgated bounds and exponents are those of Gallager [2, Sec. 5.7] and Csiszár-Körner-Marton (CKM) [10], [11] (see also [12, Ex. 10.18]). The former is based on simple inequalities such as Markov's inequality, whereas the latter is based on the type packing lemma. Gallager's approach has the notable advantage of extending to continuous channels. On the other hand, the CKM exponent can be higher when the input distribution is fixed, even though the two coincide for the optimal input distribution [13]. However, it has recently been shown in [13] that the CKM exponent can be recovered by refining Gallager's techniques.

In [4], Csiszár derived two expurgated exponents for JSCC by generalizing the techniques used in the derivation of the CKM exponent. In this paper, motivated by the preceding discussion, we provide alternative derivations of both of these exponents using a similar approach to Gallager, which can also be applied in the case of continuous channel alphabets.

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 under grant agreement 303633 and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150 and TEC2012-38800-C03-03.

Problem setup: We consider the transmission of a discrete memoryless source (DMS) over a discrete memoryless channel (DMC), described as follows. The source block length and channel block length are denoted by k and n respectively. The source, channel input and channel output alphabets are denoted by $\mathcal{V}$, $\mathcal{X}$ and $\mathcal{Y}$ respectively, and are assumed to be finite. The source $\boldsymbol{V}$ is i.i.d. on a source distribution π, i.e., $\boldsymbol{V} \sim \pi^k(\boldsymbol{v}) \triangleq \prod_{i=1}^{k} \pi(v_i)$. We assume without loss of generality that $\pi(v) > 0$ for all v. The encoder maps the source sequence $\boldsymbol{v}$ to a codeword $\boldsymbol{x}(\boldsymbol{v})$ on $\mathcal{X}^n$, and the output sequence $\boldsymbol{Y}$ is generated according to $W^n(\boldsymbol{y}|\boldsymbol{x}) \triangleq \prod_{i=1}^{n} W(y_i|x_i)$, where W is the channel transition law.

An error is said to have occurred if $\hat{\boldsymbol{v}} \neq \boldsymbol{v}$, and the average error probability for a given codebook $\mathcal{C} = \{\boldsymbol{x}(\boldsymbol{v})\}_{\boldsymbol{v}\in\mathcal{V}^k}$ is denoted by $p_{\mathrm{e}}(\mathcal{C})$. A JSCC error exponent $E(t)$ is said to be achievable at transmission ratio $t > 0$ if, for all block lengths k and n increasing in a fashion such that $\lim_{n\to\infty} \frac{k}{n} = t$, there exists a sequence of codebooks $\mathcal{C}_n$ such that $\liminf_{n\to\infty} -\frac{1}{n}\log p_{\mathrm{e}}(\mathcal{C}_n) \geq E(t)$.

Notation: Bold symbols are used for vectors (e.g., $\boldsymbol{x}$), and the corresponding i-th entry is written using a subscript (e.g., x_i). The marginals of a joint distribution P_{XY} are denoted by P_X and P_Y. The set of all empirical distributions (i.e., types [12, Ch. 2]) corresponding to length-k sequences on $\mathcal{V}$ is denoted by $\mathcal{P}_k(\mathcal{V})$. The set of all sequences of length k having a given type P_V is denoted by $T^k(P_V)$.

II. Achievable Exponents and Duality

Here we outline the existing expurgated JSCC exponents, and present equivalent expressions based on Lagrange and Fenchel duality. We define source exponent as [12, Ch. 9]

$$e(R) \triangleq \min_{P_V \,:\, H(P_V)\geq R} D(P_V\|\pi), \tag{1}$$

and we define the expurgated channel coding exponents

$$E_{\mathrm{ex}}(Q,R) \triangleq \min_{\substack{P_{X\overline{X}}:P_X=P_{\overline{X}}=Q \\ I_P(X;\overline{X})\leq R}} \mathbb{E}_P[d_B(X,\overline{X})] + I_P(X;\overline{X}) - R \tag{2}$$

$$E'_{\mathrm{ex}}(Q,R) \triangleq \min_{\substack{P_{X\overline{X}}\,:\,P_X=Q \\ I_P(X;\overline{X})\leq R}} \mathbb{E}_P[d_B(X,\overline{X})] + I_P(X;\overline{X}) - R, \tag{3}$$

where

$$d_B(x,\overline{x}) \triangleq \sum_y \sqrt{W(y|x)W(y|\overline{x})} \tag{4}$$

is the Bhattacharyya distance. Observe that $E_{\rm ex}$ is the CKM exponent [11], and that $E'_{\rm ex} \le E_{\rm ex}$ since the second marginal constraint in (2) is absent in (3).

The following achievable JSCC error exponents were given in [3], [4] for DMSs and DMCs:

$$E_{\rm J,1}(t) \triangleq \max_{Q} \min_{R\in[0,t\log|\mathcal{V}|]} te\Big(\frac{R}{t}\Big) + E_{\rm ex}(Q,R) \qquad (5)$$

$$E_{\rm J,2}(t) \triangleq \min_{R\in[0,t\log|\mathcal{V}|]} te\Big(\frac{R}{t}\Big) + \max_{Q} E'_{\rm ex}(Q,R). \qquad (6)$$

We refer to these as Csiszár's first and second exponents respectively. Observe that (5) uses the better channel coding exponent $E_{\rm ex}$ in place of $E'_{\rm ex}$, whereas (6) is better in terms of the min-max ordering. Neither exponent dominates the other in general, and the problem of determining whether the minimum and maximum can be interchanged in (5) remains open. See [4, Cor. 1-3] for conditions under which the answer is affirmative.

We proceed by providing alternative expressions for $E_{\rm J,1}$ and $E_{\rm J,2}$; we omit the proofs due to space constraints, and since the arguments are standard. We first note the following equivalent forms of (1)–(3):

$$e(R) = \sup_{\rho\ge 0} \rho R - E_{\rm s}(\rho) \qquad (7)$$

$$E_{\rm ex}(Q,R) = \sup_{\rho\ge 1} E_{\rm x}(Q,\rho) - \rho R \qquad (8)$$

$$E'_{\rm ex}(Q,R) = \sup_{\rho\ge 1} E'_{\rm x}(Q,\rho) - \rho R, \qquad (9)$$

where

$$E_{\rm s}(\rho) \triangleq \log\bigg(\sum_{v}\pi(v)^{\frac{1}{1+\rho}}\bigg)^{1+\rho} \qquad (10)$$

$$E_{\rm x}(Q,\rho) \triangleq \sup_{a(\cdot)} -\rho\sum_{x}Q(x)\log\sum_{\overline{x}}Q(\overline{x})\bigg(e^{-d_B(x,\overline{x})}\frac{e^{a(\overline{x})}}{e^{a(x)}}\bigg)^{\frac{1}{\rho}} \qquad (11)$$

$$E'_{\rm x}(Q,\rho) \triangleq \min_{Q'} -\rho\sum_{x}Q(x)\log\sum_{\overline{x}}Q'(\overline{x})e^{-\frac{d_B(x,\overline{x})}{\rho}}. \qquad (12)$$

These equivalences can be proved via Lagrange duality [14]; (7) is well-known [5], (8) was shown in [13], and (9) can be proved via the arguments used to prove (8).

We claim that the following identities hold for any DMS and DMC with zero-error capacity equal to zero:[1]

$$E_{\rm J,1}(t) = \sup_{\rho\ge 1} E_x(\rho) - tE_{\rm s}(\rho) \qquad (13)$$

$$E_{\rm J,2}(t) = \sup_{\rho\ge 1} \overline{E}'_x(\rho) - tE_{\rm s}(\rho), \qquad (14)$$

where $E_x(\rho) \triangleq \max_Q E_x(Q,\rho)$, and $\overline{E}'_x(\rho)$ is the concave hull of $E'_x(\rho) \triangleq \max_Q E'_x(Q,\rho)$. These follow by Fenchel duality using the techniques of Zhong *et al.* [5], [15]. In fact, (13) was shown in [15, Thm 5.3] in the special case that the Q maximizing $E_{\rm ex}(Q,R)$ is independent of the rate; in this case, one can safely interchange the min-max in (5) and replace $E_{\rm ex}$ and $E_{\rm x}$ by Gallager's expurgated channel coding exponent functions [2, Sec. 5.7]. The more general expression in (13) can be proved in the same way as [15, Thm 5.3] once the equivalence in (8) is established. Similarly, the ordering of the min-max in (6) translates to a concave hull operation in (14) in the same way as the case of random coding [5].

[1] We include (14) for completeness, though it will not be used in this paper.

III. Gallager-Like Derivations of $E_{\rm J,1}$ and $E_{\rm J,2}$

A. An Initial Upper Bound

We study the error probability by working with source tuples $(P_V, P_{\overline{V}})$ on a type-by-type basis, where P_V represents the actual source type, and $P_{\overline{V}}$ represents the type of an incorrect source sequence favored by the decoder. To simplify the notation, we index the source types by $i=1,\cdots,|\mathcal{P}_k(\mathcal{V})|$, and let P_i denote the corresponding i-th type. We consider a set of codeword distributions $\{P_{\boldsymbol{X}}^{(i)}\}$ depending on the source type. We let $p_{\rm e}(i,j)$ denote the probability that the transmitted sequence $\boldsymbol{v}$ has type P_i but some incorrect $\overline{\boldsymbol{v}} \in T^n(P_j)$ is chosen at the decoder (possibly with $i=j$).

Theorem 1. *Fix the codeword distributions $\{P_{\boldsymbol{X}}^{(i)}\}$ and parameters $\{\rho_{ij}\}$ such that $\rho_{ij}\ge 1$. For any non-negative function $g(\boldsymbol{v})$ depending on $\boldsymbol{v}$ only through its type, there exists a JSCC code and decoder such that*

$$p_{\rm e} \le \sum_{i=1}^{|\mathcal{P}_k(\mathcal{V})|}\sum_{j=1}^{|\mathcal{P}_k(\mathcal{V})|} p_{\rm e}(i,j), \qquad (15)$$

where, for each pair (i,j),

$$p_{\rm e}(i,j) \le \Big(4|\mathcal{P}_n(\mathcal{V})|^2\Big)^{\rho_{ij}} \times\Bigg(\sum_{\boldsymbol{v}\in T^k(P_i)}\pi^k(\boldsymbol{v})\bigg(\sum_{\overline{\boldsymbol{v}}\in T^k(P_j)}\bigg(\frac{g(\overline{\boldsymbol{v}})}{g(\boldsymbol{v})}\bigg)^{\frac{1}{\rho_{ij}}}\bigg)^{\rho_{ij}}\Bigg) \times\Bigg(\sum_{\boldsymbol{x},\overline{\boldsymbol{x}}}P_{\boldsymbol{X}}^{(i)}(\boldsymbol{x})P_{\boldsymbol{X}}^{(j)}(\overline{\boldsymbol{x}})e^{-\frac{d_B^n(\boldsymbol{x},\overline{\boldsymbol{x}})}{\rho_{ij}}}\Bigg)^{\rho_{ij}}. \qquad (16)$$

Before proving this result (see Section III-D), we show how it can be used to derive $E_{\rm J,1}$ and $E_{\rm J,2}$.

B. Derivation of $E_{\rm J,1}$ Using Theorem 1

We set $g(\boldsymbol{v}) = \pi^k(\boldsymbol{v})^s$ for some $s\in[0,1]$, let each $P_{\boldsymbol{X}}^{(i)}(\boldsymbol{x})$ be a common distribution $P_{\boldsymbol{X}}(\boldsymbol{x})$ to be specified shortly, and let each ρ_{ij} equal a common value ρ. Hence, (16) yields

$$p_{\rm e}(i,j) \le \Big(4|\mathcal{P}_n(\mathcal{V})|^2\Big)^{\rho} \times\Bigg(\sum_{\boldsymbol{v}}\pi^k(\boldsymbol{v})\bigg(\sum_{\overline{\boldsymbol{v}}}\bigg(\frac{\pi^k(\overline{\boldsymbol{v}})}{\pi^k(\boldsymbol{v})}\bigg)^{\frac{s}{\rho}}\bigg)^{\rho}\Bigg) \times\Bigg(\sum_{\boldsymbol{x},\overline{\boldsymbol{x}}}P_{\boldsymbol{X}}(\boldsymbol{x})P_{\boldsymbol{X}}(\overline{\boldsymbol{x}})e^{-\frac{d_B^n(\boldsymbol{x},\overline{\boldsymbol{x}})}{\rho}}\Bigg)^{\rho}, \qquad (17)$$

where we have upper bounded the summations over $\boldsymbol{v}$ and $\overline{\boldsymbol{v}}$ by expanding them to all sequences.

Observe that the bound in (17) contains a source factor and a channel factor. With $s = \frac{\rho}{1+\rho}$, the source factor equals

$$\sum_{\boldsymbol{v}} \pi^k(\boldsymbol{v}) \Bigg(\sum_{\overline{\boldsymbol{v}}} \bigg(\frac{\pi^k(\overline{\boldsymbol{v}})}{\pi^k(\boldsymbol{v})} \bigg)^{\frac{1}{1+\rho}} \Bigg)^{\rho} \tag{18}$$

$$= \sum_{\boldsymbol{v}} \pi^k(\boldsymbol{v})^{1-\frac{\rho}{1+\rho}} \Bigg(\sum_{\overline{\boldsymbol{v}}} \pi^k(\overline{\boldsymbol{v}})^{\frac{1}{1+\rho}} \Bigg)^{\rho} \tag{19}$$

$$= \Bigg(\sum_{\boldsymbol{v}} \pi^k(\boldsymbol{v})^{\frac{1}{1+\rho}} \Bigg)^{1+\rho} = e^{kE_{\mathrm{s}}(\rho)}, \tag{20}$$

where the last step follows from (10).

For the channel factor, we consider cost-constrained random coding [16]; while constant-composition random coding also suffices, we focus on the former since it extends immediately to channels with continuous input alphabets upon suitably replacing summations by integrals. The codeword distribution is given by

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\mu_n} \prod_{i=1}^{n} Q(x_i) \mathbb{1}\{\boldsymbol{x} \in \mathcal{D}_n\}, \tag{21}$$

where

$$\mathcal{D}_n \triangleq \bigg\{ \boldsymbol{x} \,:\, \bigg| \frac{1}{n} \sum_{i=1}^{n} a_l(x_i) - \phi_l \bigg| \le \delta,\ l = 1, \dots, L \bigg\}, \tag{22}$$

and where μ_n is a normalizing constant, δ is a positive constant (independent of n), and for each $l = 1, \dots, L$, $a_l(\cdot)$ is an auxiliary cost function and $\phi_l \triangleq \mathbb{E}_Q[a_l(X)]$ is its mean. Roughly speaking, each codeword is generated according to an i.i.d. distribution conditioned on the empirical mean of each $a_l(x)$ being close to the true mean. We have the following.

Proposition 2. *Fix an input alphabet $\mathcal{X}$, an input distribution $Q \in \mathcal{P}(\mathcal{X})$, and auxiliary cost functions $a_1(\cdot), \dots, a_L(\cdot)$ such that $|a_l(x)|$ is uniformly bounded in $l = 1, \dots, L$ and $x \in \mathcal{X}$. For any $\delta > 0$ in* (22)*, there exists $\eta > 0$ such that the normalizing constant in* (21) *satisfies $\mu_n \ge 1 - 2Le^{-n\eta}$.*

Proof: For any $l = 1, \dots, L$, since $|a_l(x)|$ is bounded, we have from Hoeffding's inequality that $\big| \frac{1}{n} \sum_{i=1}^{n} a_l(x_i) - \phi_l \big| \le \delta$ with probability at least $1 - 2e^{-n\eta}$. The proposition follows by taking the union bound over the L cost functions. ∎

We now set $L = 2$, let $a_1(\cdot)$ be arbitrary for now, and choose $a_2(x) = \log \sum_{\overline{x}} Q(\overline{x}) \big(e^{-d_B(x,\overline{x})} \frac{e^{a_1(\overline{x})}}{e^{a_1(x)}} \big)^{\frac{1}{\rho}}$. Letting $a_l^n(\boldsymbol{x}) \triangleq \sum_{i=1}^{n} a_l(x_i)$ and $Q^n(\boldsymbol{x}) \triangleq \prod_{i=1}^{n} Q(x_i)$, we recall the following steps from [13]:

$$\sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}(\boldsymbol{x}) P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}) e^{-\frac{d_B^n(\boldsymbol{x},\overline{\boldsymbol{x}})}{\rho}}$$

$$\le e^{\frac{2\delta n}{\rho}} \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}(\boldsymbol{x}) P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}) \Bigg(e^{-d_B^n(\boldsymbol{x},\overline{\boldsymbol{x}})} \frac{e^{a_1^n(\overline{\boldsymbol{x}})}}{e^{a_1^n(\boldsymbol{x})}} \Bigg)^{\frac{1}{\rho}} \tag{23}$$

$$\le \frac{e^{\frac{2\delta n}{\rho}}}{\mu_n} \max_{\boldsymbol{x} \in \mathcal{D}_n} \sum_{\overline{\boldsymbol{x}}} Q^n(\overline{\boldsymbol{x}}) \Bigg(e^{-d_B^n(\boldsymbol{x},\overline{\boldsymbol{x}})} \frac{e^{a_1^n(\overline{\boldsymbol{x}})}}{e^{a_1^n(\boldsymbol{x})}} \Bigg)^{\frac{1}{\rho}} \tag{24}$$

$$= \frac{e^{\frac{2\delta n}{\rho}}}{\mu_n} \exp\Big(\max_{\boldsymbol{x} \in \mathcal{D}_n} a_2^n(\boldsymbol{x}) \Big) \tag{25}$$

$$\le \frac{e^{\frac{2\delta n}{\rho} + \delta n}}{\mu_n} \exp\big(n\mathbb{E}_Q[a_2(X)] \big) \tag{26}$$

where (23) follows since $a_1^n(\cdot)$ is δn-close to $n\phi_1$ by (22), (24) follows from (21), (25) follows from the choice of $a_2(\cdot)$, and (26) follows since $a_2^n(\boldsymbol{x})$ is δn-close to $n\mathbb{E}_Q[a_2(X)]$.

Substituting (26) and the choice of $a_2(\cdot)$ into (17) and renaming $a_1(\cdot)$ as $a(\cdot)$, we see that the channel factor is upper bounded by $\frac{e^{(2+\rho)\delta n}}{\mu_n^{\rho}} e^{-nE_{\mathrm{x}}(Q,\rho)}$ (see (11)). Hence, using (20), Proposition 2, the fact that δ can be arbitrarily small, and $|\mathcal{P}_k(\mathcal{V})| \le (n+1)^{|\mathcal{V}|-1}$ [12, Ch. 2], we obtain the exponent $E_{\mathrm{J},1}$ in the form given in (13) upon optimizing ρ, $a(\cdot)$ and Q.

C. Derivation of $E_{\mathrm{J},2}$ Using Theorem 1

The expression in (16) can be simplified by noting that both $g(\cdot)$ and $\pi^k(\cdot)$ take fixed values among the sequences within a given type class (for $g(\cdot)$, this is an assumption of the theorem). Specifically, fixing the sequences $\boldsymbol{v} \in T^k(P_i)$ and $\overline{\boldsymbol{v}} \in T^k(P_j)$ arbitrarily, (16) can be written as

$$p_{\mathrm{e}}(i,j) \le \Big(4|\mathcal{P}_n(\mathcal{V})|^2 \Big)^{\rho_{ij}} \Bigg(\big|T^k(P_i)\big| \pi^k(\boldsymbol{v}) \frac{g(\overline{\boldsymbol{v}})}{g(\boldsymbol{v})} \Bigg) \times \Bigg(\big|T^k(P_j)\big| \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}^{(i)}(\boldsymbol{x}) P_{\boldsymbol{X}}^{(j)}(\overline{\boldsymbol{x}}) e^{-\frac{d_B^n(\boldsymbol{x},\overline{\boldsymbol{x}})}{\rho_{ij}}} \Bigg)^{\rho_{ij}}, \tag{27}$$

where we moved $\big|T^k(P_j)\big|$ from the source factor to the channel factor for later convenience. We choose $g(\cdot)$ as follows:

$$g(\boldsymbol{v}) = \big|T^k(P_i)\big| \pi^k(\boldsymbol{v}) \tag{28}$$

$$g(\overline{\boldsymbol{v}}) = \big|T^k(P_j)\big| \pi^k(\overline{\boldsymbol{v}}). \tag{29}$$

We now fix the distributions $\{Q_i\}$ and choose $P_{\boldsymbol{X}}^{(i)}$ to be the cost-constrained distribution (21) with Q_i in place of Q, and with $L = |\mathcal{P}_k(\mathcal{V})|$ auxiliary costs[2] of the form $a_l(x) = \log \sum_{\overline{x}} Q_l(\overline{x}) e^{-d_B(x,\overline{x})/\rho}$. Thus, L is polynomial in k, meaning that the prefactor μ_n still tends to one exponentially fast by Proposition 2.

Substituting (28)–(29) into (27), we obtain

$$\begin{aligned} &p_{\mathrm{e}}(i,j) \\ &\le \Big(4|\mathcal{P}_n(\mathcal{V})|^2 \Big)^{\rho_{ij}} \Big(\big|T^k(P_j)\big| \pi^k(\overline{\boldsymbol{v}}) \Big) \\ &\quad \times \Bigg(\big|T^k(P_j)\big| \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}^{(i)}(\boldsymbol{x}) P_{\boldsymbol{X}}^{(j)}(\overline{\boldsymbol{x}}) e^{-\frac{d_B^n(\boldsymbol{x},\overline{\boldsymbol{x}})}{\rho_{ij}}} \Bigg)^{\rho_{ij}} \qquad (30) \\ &\le \Big(4|\mathcal{P}_n(\mathcal{V})|^2 \Big)^{\rho_{ij}} \Big(\big|T^k(P_j)\big| \pi^k(\overline{\boldsymbol{v}}) \Big) \\ &\quad \times \max_{P_{\boldsymbol{X}}} \Bigg(\big|T^k(P_j)\big| \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}(\boldsymbol{x}) P_{\boldsymbol{X}}^{(j)}(\overline{\boldsymbol{x}}) e^{-\frac{d_B^n(\boldsymbol{x},\overline{\boldsymbol{x}})}{\rho_{ij}}} \Bigg)^{\rho_{ij}}, \qquad (31) \end{aligned}$$

[2] It suffices to have $L = |\mathcal{X}| - 1$ auxiliary costs in the finite-alphabet setting, but we provide a more general argument here since it extends to channels with continuous alphabets upon suitably replacing summations by integrals, under mild technical assumptions ensuring the exponential decay in Proposition 2 (e.g., boundedness or sub-Gaussianity of the auxiliary costs).

where the maximum over $P_{\boldsymbol{X}}$ is over the set of $|\mathcal{P}_k(\mathcal{V})|$ cost-constrained distributions characterized by the $\{Q_i\}$. Since d_B is symmetric in its arguments, this yields

$$p_{\mathrm{e}}(i,j) \le \Big(4|\mathcal{P}_n(\mathcal{V})|^2\Big)^{\rho_{ij}} \Big(\big|T^k(P_j)\big|\pi^k(\overline{\boldsymbol{v}})\Big) \times \max_{P_{\boldsymbol{X}}} \left(\big|T^k(P_j)\big| \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}^{(j)}(\boldsymbol{x}) P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}) e^{-\frac{d_B^n(\boldsymbol{x},\overline{\boldsymbol{x}})}{\rho_{ij}}}\right)^{\rho_{ij}}. \quad (32)$$

We proceed by upper bounding the source and channel factors in (32). For the former, we combine the simple identity $\pi^k(\overline{\boldsymbol{v}}) = e^{k\mathbb{E}_{P_j}[\log \pi(V)]}$ (which holds since $\overline{\boldsymbol{v}} \in T^k(P_j)$) with the inequality $\big|T^k(P_j)\big| \le e^{kH(P_j)}$ [12, Ch. 2] to obtain

$$\big|T^k(P_j)\big|\pi^k(\overline{\boldsymbol{v}}) \le e^{kD(P_j\|\pi)}. \quad (33)$$

Recalling that we chose the cost-constrained distribution in (21) to contain all auxiliary costs of the form $a_l(x) = \log \sum_{\overline{x}} Q_l(\overline{x}) e^{-d_B(x,\overline{x})/\rho}$ for $l = 1, \ldots, |\mathcal{P}_k(\mathcal{V})|$, we obtain via similar steps to (23)–(26) that the channel factor in (32) is upper bounded by a subexponential factor times

$$\Big(\big|T^k(P_j)\big| e^{-nE'_{\mathrm{x}}(Q_j,\rho_{ij})}\Big)^{\rho_{ij}}, \quad (34)$$

where E'_{x} is defined in (12).

Substituting (33) and (34) into (32), optimizing ρ_{ij} and $\{Q_j\}$, and again using the fact that $\big|T^k(P_j)\big| \le e^{kH(P_j)}$, we see that the error exponent corresponding to the right-hand side of (32) is lower bounded by

$$E_{\mathrm{J},2}^{(j)}(t) \triangleq tD(P_j\|\pi) + \max_{Q_j} E'_{\mathrm{ex}}(Q_j, tH(P_j)) \quad (35)$$

$$\ge \min_{P_V : H(P_V) \ge H(P_j)} tD(P_V\|\pi) + \max_{Q_j} E'_{\mathrm{ex}}(Q_j, tH(P_j)) \quad (36)$$

$$= te(H(P_j)) + \max_{Q} E'_{\mathrm{ex}}(Q, tH(P_j)), \quad (37)$$

where the final step uses (1) and renames Q_j as Q. The overall exponent corresponding to p_{e} in (15) is thus bounded by

$$\min_j E_{\mathrm{J},2}^{(j)}(t) \ge \min_j te(H(P_j)) + \max_Q E'_{\mathrm{ex}}(Q, tH(P_j)) \quad (38)$$

$$\ge \min_{R\in[0,\log|\mathcal{V}|]} te(R) + \max_Q E'_{\mathrm{ex}}(Q, tR) \quad (39)$$

$$= \min_{R\in[0,t\log|\mathcal{V}|]} te\Big(\frac{R}{t}\Big) + \max_Q E'_{\mathrm{ex}}(Q, R), \quad (40)$$

where (39) follows since $H(P_j)$ is always between 0 and $\log|\mathcal{V}|$, and (40) follows from a change of variable. We have thus recovered $E_{\mathrm{J},2}$ as given in (6).

D. Proof of Theorem 1

The proof is based on a decoding rule of the form

$$\hat{\boldsymbol{v}} = \arg\max_{\boldsymbol{v}} g(\boldsymbol{v})^2 W^n(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{v})), \quad (41)$$

where $g(\boldsymbol{v})$ is given in the theorem statement. We assume that ties are broken as errors.

We consider a modified JSCC setup with duplicates of each source sequence, and then apply expurgation on a type-by-type basis in such a way that there are no duplicates. The modified setup is as follows: A random variable Z takes values uniformly on a finite alphabet $\mathcal{Z}$. Each pair $(\boldsymbol{v}, z)$ is assigned a codeword $\boldsymbol{x}(\boldsymbol{v}, z)$, and the decoder estimates

$$(\hat{\boldsymbol{v}}, \hat{z}) = \arg\max_{(\boldsymbol{v},z)} g(\boldsymbol{v})^2 W^n(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{v}, z)). \quad (42)$$

An error occurs if $(\hat{\boldsymbol{v}}, \hat{z}) \ne (\boldsymbol{v}, z)$ (or if there is a tie), and the error probability given that $(\boldsymbol{V}, Z) = (\boldsymbol{v}, z)$ is denoted by $\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, \mathsf{c})$, where $\mathsf{c} = \{\boldsymbol{x}(\boldsymbol{v}, z)\}$ is the codebook. Hence,

$$\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, \mathsf{c}) \le \sum_j \tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, j, \mathsf{c}), \quad (43)$$

where $\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, j, \mathsf{c})$ is the probability given $(\boldsymbol{V}, Z) = (\boldsymbol{v}, z)$ that some $(\overline{\boldsymbol{v}}, \overline{z})$ with $\overline{\boldsymbol{v}} \in T^k(P_j)$ yields a decoding metric at least as high as that of $(\boldsymbol{v}, z)$.

Consider a random codebook C in which each codeword corresponding to a pair $(\boldsymbol{v}, z)$ with $\boldsymbol{v} \in T^k(P_i)$ is generated according to $P_{\boldsymbol{X}}^{(i)}$ (i.e., the codeword distribution depends on the source type). Hence, $\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, \mathsf{C})$ is a random variable. By Markov's inequality and the union bound, we have for any $\rho_{ij} > 0$, $\eta > 0$ and pair $(\boldsymbol{v}, z)$ that

$$\mathbb{P}\bigg[\bigcup_j \Big\{\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, j, \mathsf{C})^{\frac{1}{\rho_{ij}}} \ge (1+\eta)\mathbb{E}\big[\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, j, \mathsf{C})^{\frac{1}{\rho_{ij}}}\big]\Big\}\bigg] \le \frac{|\mathcal{P}_n(\mathcal{V})|}{1+\eta}. \quad (44)$$

For each source type $P_V \in \mathcal{P}_k(\mathcal{V})$, let $N(P_V)$ be the number of $(\boldsymbol{v}, z)$ pairs with $\boldsymbol{v} \in T^n(P_V)$, and let $N_0(P_V, \mathsf{C})$ be the random number of such pairs that fail to satisfy the event in (44), i.e., the number of $(\boldsymbol{v}, z)$ with $\boldsymbol{v} \in T^n(P_V)$ such that

$$\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, j, \mathsf{c})^{\frac{1}{\rho_{ij}}} < (1+\eta)\mathbb{E}\big[\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, j, \mathsf{C})^{\frac{1}{\rho_{ij}}}\big], \quad \forall j. \quad (45)$$

We have from (44) that

$$\mathbb{E}\big[N_0(P_V, \mathsf{C})\big] \ge N(P_V)\Big(1 - \frac{|\mathcal{P}_n(\mathcal{V})|}{1+\eta}\Big). \quad (46)$$

Since $N_0(P_V, \mathsf{C}) \le N(P_V)$ with probability one, the reverse Markov inequality[3] gives for any $\alpha < 1 - \frac{1}{1+\eta}$ that

$$\mathbb{P}\big[N_0(P_V, \mathsf{C}) > \alpha N(P_V)\big] \ge \frac{1 - \frac{|\mathcal{P}_n(\mathcal{V})|}{1+\eta} - \alpha}{1-\alpha}, \quad (47)$$

or equivalently

$$\mathbb{P}\big[N_0(P_V, \mathsf{C}) \le \alpha N(P_V)\big] \le \frac{\frac{|\mathcal{P}_n(\mathcal{V})|}{1+\eta}}{1-\alpha} \quad (48)$$

$$= \frac{|\mathcal{P}_n(\mathcal{V})|}{(1+\eta)(1-\alpha)}. \quad (49)$$

Thus, using the union bound, we obtain

$$\mathbb{P}\bigg[\bigcup_{P_V\in\mathcal{P}_k(\mathcal{V})} \Big\{N_0(P_V, \mathsf{C}) \le \alpha N(P_V)\Big\}\bigg] \le \frac{|\mathcal{P}_k(\mathcal{V})|^2}{(1+\eta)(1-\alpha)}. \quad (50)$$

[3]If $A \in [0, a_{\max}]$ with probability one, then $\mathbb{P}[A > a] \ge \frac{\mathbb{E}[A]-a}{a_{\max}-a}$ for any $a < \mathbb{E}[X]$. This is proved by applying Markov's inequality to the non-negative random variable $a_{\max} - A$.

Fixing $\gamma > 0$ and choosing $\alpha = \frac{1}{2}$, $|\mathcal{Z}| = 2$, and $\eta = 2|\mathcal{P}_n(\mathcal{V})|^2+\gamma-1$, we see that the right-hand side of (50) equals $\frac{2|\mathcal{P}_k(\mathcal{V})|}{2|\mathcal{P}_k(\mathcal{V})|+\gamma} < 1$, and we conclude that there exists at least one codebook c satisfying $N_0(P_V, \mathsf{c}) > \alpha N(P_V)$ for every type P_V. Since $\alpha = \frac{1}{2}$, this means that at least half of the $(\boldsymbol{v}, z)$ pairs corresponding to each source type P_V satisfy (45). Let $\mathcal{C}(P_V)$ denote any such collection of $\frac{1}{2}N(P_V)$ codewords.

We now construct a *standard* JSCC code from $\{\mathcal{C}(P_V)\}$. Since $|\mathcal{Z}| = 2$, we see that $\frac{1}{2}N(P_V)$ is precisely the number of $\boldsymbol{v}$ sequences having type P_V. Let $\mathcal{C}$ be a code for which, for each P_V, the sequences $\boldsymbol{v} \in T^n(P_V)$ are assigned from $\mathcal{C}(P_V)$ in an arbitrary one-to-one fashion. Since the decoding rule (see (41)) depends on $\boldsymbol{v}$ only through its type, any such assignment of codewords yields the same error probability.

For a fixed sequence $\boldsymbol{v}$, consider the codeword $\boldsymbol{x}$ from the preceding construction, and let $(\boldsymbol{v}', z')$ be the pair corresponding to that codeword in the modified setup. We see that $p_{\mathrm{e}}(\boldsymbol{v}, \mathcal{C}) \le \tilde{p}_{\mathrm{e}}(\boldsymbol{v}', z', \mathsf{c})$ by noting that $\boldsymbol{v}$ and $\boldsymbol{v}'$ have the same type by construction, the decoding rules for the two setups coincide, and $\mathcal{C}$ was obtained from c by removing codewords. Letting $p_{\mathrm{e}}(\boldsymbol{v}, j, \mathcal{C})$ be the probability (for the standard JSCC setup) given $\boldsymbol{V} = \boldsymbol{v}$ that some $\overline{\boldsymbol{v}} \in T^k(P_j)$ yields a decoding metric at least as high as that of $\boldsymbol{v}$, we conclude from (45) that there exists a codebook $\mathcal{C}$ such that

$$p_{\mathrm{e}}(\boldsymbol{v}, j, \mathcal{C}) < \Big(\big(2|\mathcal{P}_k(\mathcal{V})|^2+\gamma\big)\mathbb{E}\big[\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, j, \mathsf{C})^{\frac{1}{\rho_{ij}}}\big]\Big)^{\rho_{ij}} \quad (51)$$

for (i, j) and $\boldsymbol{v} \in T^k(P_i)$ (recall $\eta = 2|\mathcal{P}_k(\mathcal{V})| + \gamma - 1$).

We can upper bound the inner expectation in (51) using standard arguments [2], [13]. Letting $\boldsymbol{X}(\boldsymbol{v}, z)$ denote the random codeword associated with $(\boldsymbol{v}, z)$ in the above modified JSCC setup with $\mathcal{Z} = \{1, 2\}$, we have

$$\begin{aligned}
&\mathbb{E}\big[\tilde{p}_{\mathrm{e}}(\boldsymbol{v}, z, j, , \mathsf{C})^{\frac{1}{\rho_{ij}}}\big] \\
&= \mathbb{E}\Bigg[\mathbb{P}\Bigg[\bigcup_{(\overline{\boldsymbol{v}},\overline{z})\neq(\boldsymbol{v},z),\, \overline{\boldsymbol{v}}\in T^k(P_j)} \Big\{g(\overline{\boldsymbol{v}})^2 W^n(\boldsymbol{Y}|\boldsymbol{X}(\overline{\boldsymbol{v}}, \overline{z})) \\
&\qquad \ge g(\boldsymbol{v})^{2s} W^n(\boldsymbol{Y}|\boldsymbol{X}(\boldsymbol{v}, z))\Big\} \,\Big|\, \mathsf{C}\Bigg]^{\frac{1}{\rho_{ij}}}\Bigg] && (52) \\
&\le \mathbb{E}\Bigg[\Bigg(\sum_{(\overline{\boldsymbol{v}},\overline{z})\neq(\boldsymbol{v},z),\, \overline{\boldsymbol{v}}\in T^k(P_j)} \mathbb{P}\Big[g(\overline{\boldsymbol{v}})^2 W^n(\boldsymbol{Y}|\boldsymbol{X}(\overline{\boldsymbol{v}}, \overline{z})) \\
&\qquad \ge g(\boldsymbol{v})^{2s} W^n(\boldsymbol{Y}|\boldsymbol{X}(\boldsymbol{v}, z)) \,\Big|\, \mathsf{C}\Big]\Bigg)^{\frac{1}{\rho_{ij}}}\Bigg] && (53) \\
&\le \sum_{\overline{\boldsymbol{v}}\in T^k(P_j)} \sum_{z=1}^{2} \mathbb{E}\Bigg[\Bigg(\mathbb{P}\Big[g(\overline{\boldsymbol{v}})^2 W^n(\boldsymbol{Y}|\overline{\boldsymbol{X}}(\overline{\boldsymbol{v}}, \overline{z})) \\
&\qquad \ge g(\boldsymbol{v})^{2s} W^n(\boldsymbol{Y}|\boldsymbol{X}(\boldsymbol{v}, z)) \,\Big|\, \mathsf{C}\Big]\Bigg)^{\frac{1}{\rho_{ij}}}\Bigg] && (54) \\
&\le 2 \sum_{\overline{\boldsymbol{v}}\in T^k(P_j)} \sum_{\boldsymbol{x},\overline{\boldsymbol{x}}} P_{\boldsymbol{X}}^{(i)}(\boldsymbol{x}) P_{\boldsymbol{X}}^{(j)}(\overline{\boldsymbol{x}}) \\
&\qquad \times \left(\frac{g(\overline{\boldsymbol{v}})}{g(\boldsymbol{v})} \sum_{\boldsymbol{y}} \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x}) W^n(\boldsymbol{y}|\overline{\boldsymbol{x}})}\right)^{\frac{1}{\rho_{ij}}}, && (55)
\end{aligned}$$

where (52) follows from the decoding rule in (42), (53) follows from the union bound, (54) follows from the inequality $\big(\sum_i a_i\big)^{\frac{1}{\rho}} \le \sum_i a_i^{\frac{1}{\rho}}$ for $\rho \ge 1$, and (55) follows from Markov's inequality as per Gallager's analysis [2]. Substituting (55) into (51), applying the identity

$$p_{\mathrm{e}}(i, j) = \sum_{\boldsymbol{v}\in T^k(P_i)} \pi^k(\boldsymbol{v}) p_{\mathrm{e}}(\boldsymbol{v}, j, \mathcal{C}), \quad (56)$$

and performing simple rearrangements, we obtain (16) with $2(2|\mathcal{P}_k(\mathcal{V})|^2 + \gamma)$ in place of $4|\mathcal{P}_k(\mathcal{V})|^2$. Since γ can be arbitrarily small, this completes the proof.

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] I. Csiszár, "Joint source-channel error exponent," *Prob. Contr. Inf. Theory*, vol. 9, pp. 315–328, 1980.

[4] ——, "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.

[5] Y. Zhong, F. Alajaji, and L. Campbell, "On the joint source-channel coding error exponent for discrete memoryless systems," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1450–1468, April 2006.

[6] A. Tauste Campo, G. Vazquez-Vilar, A. Guillén i Fàbregas, T. Koch, and A. Martinez, "A derivation of the source-channel error exponent using nonidentical product distributions," *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3209–3217, June 2014.

[7] D. Wang, A. Ingber, and Y. Kochman, "The dispersion of joint source-channel coding," 2011, http://arxiv.org/abs/1109.6310.

[8] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2545–2575, May 2013.

[9] A. Tauste Campo, G. Vazquez-Vilar, A. Guillén i Fàbregas, and A. Martinez, "Random-coding joint source-channel bounds," in *IEEE Int. Symp. Inf. Theory*, St. Petersburg, 2011.

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

[11] 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.

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

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

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

[15] Y. Zhong, "Joint source-channel coding reliability function for single and multi-terminal communication systems," Ph.D. dissertation, Queen's University, 2009, http://www.mast.queensu.ca/~web/Papers/yangfan-phd08.pdf.

[16] 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.