Dual-Domain Exponent of Maximum Mutual Information Decoding

Seyed AmirPouya Moeini
University of Cambridge
sam297@cam.ac.uk

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

Abstract—**This paper provides a dual domain derivation of the error exponent of maximum mutual information (MMI) decoding with constant composition codes, showing it coincides with that of maximum likelihood decoding for discrete memoryless channels. The analysis is further extended to joint source-channel coding, demonstrating that the generalized MMI decoder achieves the same random coding error exponent as the maximum a posteriori decoder.**

I. INTRODUCTION

The random coding error exponent is defined as the asymptotic limit of the negative normalized logarithm of the expected error probability, given by

$$E_{\mathrm{r}}(R) \triangleq \lim_{n\to\infty} -\frac{1}{n} \log \bar{p}_e, \qquad (1)$$

where R is the coding rate, and $\bar{p}_e$ denotes the average error probability over all codebooks in the ensemble.

When messages are equiprobable, Csiszár and Körner [1, Theorem 10.2] (see also Gallager [2, Section 6]) showed that, for discrete memoryless channels (DMCs) and under constant-composition coding, the maximum mutual information (MMI) decoder achieves the same random coding error exponent as the maximum likelihood (ML) decoder. The MMI decoder is a suboptimal mismatched decoder that does not know the channel transition probability and yet attains the same error exponent as ML decoding. This notion of optimality with respect to the random coding error exponent is commonly termed as *universality*.

Csiszár and Körner's expression results in the minimization over joint distributions subject to specific constraints. These expressions are typically referred to as primal domain, and are typically obtained using the method of types [1, Chapter 2]. Instead, dual expressions are commonly formulated as maximizations over auxiliary parameters, with key approaches for deriving random coding bounds in the dual domain being those of Gallager [3, Section 5.6] and Poltyrev [4]. A dual-domain derivation offers notable advantages, such as enabling achievable exponents to be obtained for any parameter choice and providing a natural extension to arbitrary alphabets.

Existing results of the MMI decoder have been limited to the primal domain, as its multi-letter nature makes a dual-domain analysis difficult. In this work, we revisit the universality of the MMI decoder in channel coding by deriving its random coding error exponent in the dual domain. We then extend our analysis to joint source-channel coding (JSCC), where messages are not equiprobable, by addressing the generalized MMI decoder introduced by Csiszár [5]. To the best of our knowledge, such dual-domain derivations for the MMI decoder have not been previously appeared in the literature.

II. NOTATIONS

In this paper, scalar random variables are denoted by uppercase letters, their sample values by lowercase letters, and their alphabets by calligraphic letters. Random vectors are represented in boldface. 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 $\limsup_{n\to\infty} \frac{1}{n} \log \frac{f_n}{g_n} \leq 0$.

The type of a sequence $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ is its empirical distribution, defined by

$$\hat{P}_{\boldsymbol{x}}(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = x\}. \qquad (2)$$

Similarly, the joint type and conditional types of a pair of sequences are denoted by $\hat{P}_{\boldsymbol{xy}}$, $\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}$, and $\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}$, and are defined in an analogous manner. Throughout this paper, we use $\hat{P}_{\boldsymbol{x}}(\boldsymbol{x})$ for convenience to represent $\hat{P}_{\boldsymbol{x}}(\boldsymbol{x}) = \prod_{i=1}^{n} \hat{P}_{\boldsymbol{x}}(x_i)$. Observe that $\hat{P}_{\boldsymbol{x}}(\boldsymbol{x}) = \exp(-nH(\hat{P}_{\boldsymbol{x}}))$, where $H(P)$ represents the entropy of the distribution P. The set of all probability distributions on an alphabet $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$, while $\mathcal{P}_n(\mathcal{X})$ represents the set of empirical distributions (types) for vectors in $\mathcal{X}^n$. It is shown in [1, Lemma 2.2] that the total number of types grows polynomially with n, which implies $|\mathcal{P}_n(\mathcal{X})| \doteq 1$. For $P_X \in \mathcal{P}_n(\mathcal{X})$, the type class $\mathcal{T}^n(P_X)$ consists of all sequences in $\mathcal{X}^n$ with type P_X. It is shown in [1, Lemma 2.3] that $|\mathcal{T}^n(P_X)| \doteq \exp(nH(P_X))$. Under the constant-composition random coding ensemble with input distribution Q_X, the codewords are independently drawn from the codeword distribution

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{|\mathcal{T}^n(Q_X)|} \mathbb{1}\{\boldsymbol{x} \in \mathcal{T}^n(Q_X)\}. \qquad (3)$$

III. CHANNEL CODING

We study communication over a DMC W. A codebook $\mathcal{C}_n$ consists of M codewords $\mathcal{C}_n = \{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(M)\}$, where each codeword $\boldsymbol{x}(m) \in \mathcal{X}^n$ for $m \in \{1, \ldots, M\}$. The code rate R is defined as $R = \frac{1}{n}\log M$, measured in nats per channel use. A message m is selected equiprobably from $\{1, \ldots, M\}$, and the corresponding codeword $\boldsymbol{x}(m)$ is sent through the channel. The channel produces an output sequence $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ based on the conditional probability $W^n(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n W(y_i|x_i)$. In [1, Chapter 10], Csiszár and Körner used the method of types [1, Chapter 2] to show that the MMI decoder achieves the same random coding error exponent as the ML decoder. In this section, we redrive this result in the dual domain.

Given the channel output $\boldsymbol{y}$, the MMI decoder selects the message $\hat{m}$ that maximizes the mutual information induced by the joint type between $\boldsymbol{y}$ and each of the codewords,

$$\hat{m} = \underset{m \in \{1, \ldots, M\}}{\operatorname{argmax}} I(\hat{P}_{\boldsymbol{x}(m)\boldsymbol{y}}), \qquad (4)$$

where $\hat{P}_{\boldsymbol{x}(m)\boldsymbol{y}}$ denotes the joint type of $\boldsymbol{x}(m)$ and $\boldsymbol{y}$. This decoder is equivalent to selecting the message that maximizes the conditional type. To see this, observe that

$$\underset{m \in \{1, \ldots, M\}}{\operatorname{argmax}} I(\hat{P}_{\boldsymbol{x}(m)\boldsymbol{y}})$$

$$= \underset{m \in \{1, \ldots, M\}}{\operatorname{argmax}} \exp\left(nI(\hat{P}_{\boldsymbol{x}(m)\boldsymbol{y}})\right) \qquad (5)$$

$$= \underset{m \in \{1, \ldots, M\}}{\operatorname{argmax}} \exp\left(n\left[H(\hat{P}_{\boldsymbol{y}}) - H(\hat{P}_{\boldsymbol{y}|\boldsymbol{x}(m)})\right]\right) \quad (6)$$

$$= \underset{m \in \{1, \ldots, M\}}{\operatorname{argmax}} \exp\left(-nH(\hat{P}_{\boldsymbol{y}|\boldsymbol{x}(m)})\right). \qquad (7)$$

The equality in (7) follows from the fact that $H(\hat{P}_{\boldsymbol{y}})$ acts as a constant here, as it does not depend on m. Furthermore, the conditional entropy $H(\hat{P}_{\boldsymbol{y}|\boldsymbol{x}(m)})$ can be expressed as

$$-nH(\hat{P}_{\boldsymbol{y}|\boldsymbol{x}(m)}) = \sum_{(a,b)} n\hat{P}_{\boldsymbol{x}(m)\boldsymbol{y}}(a,b)\log \hat{P}_{\boldsymbol{y}|\boldsymbol{x}(m)}(b|a) \quad (8)$$

$$= \log \prod_{(a,b)} \left[\hat{P}_{\boldsymbol{y}|\boldsymbol{x}(m)}(b|a)\right]^{n\hat{P}_{\boldsymbol{x}(m)\boldsymbol{y}}(a,b)} \qquad (9)$$

$$= \log \prod_{i=1}^n \hat{P}_{\boldsymbol{y}|\boldsymbol{x}(m)}(y_i|x_{m,i}). \qquad (10)$$

Thus, we conclude that this decoder selects the message $\hat{m}$ as

$$\hat{m} = \underset{m \in \{1, \ldots, M\}}{\operatorname{argmax}} \prod_{i=1}^n \hat{P}_{\boldsymbol{y}|\boldsymbol{x}(m)}(y_i|x_{m,i}). \qquad (11)$$

This form of the MMI decoder is more amenable for dual-domain derivations.

Theorem 1. *For a DMC W and rate R, under constant-composition coding with Q_X, we have*

$$\bar{p}_e^{\mathrm{mmi}} \dot{\leq} \min_{0 \leq \rho \leq 1} \exp\left(-n\left[E_0^{\mathrm{ml}}(Q_X, \rho) - \rho R\right]\right), \qquad (12)$$

where

$$E_0^{\mathrm{ml}}(Q_X, \rho)$$

$$= \sup_{r(.)} -\log \sum_y \left(\sum_x Q_X(x)W(y|x)^{\frac{1}{1+\rho}} e^{r(x)-\phi_r}\right)^{1+\rho} \quad (13)$$

and $\phi_r = \sum_x Q_X(x)r(x)$.

Proof. We begin with the random coding union (RCU) bound, presented in [6, Theorem 1], which states that for a generic decoding metric q, we have

$$\bar{p}_e^q \dot{=} \mathbb{E}\left[\min\left\{1, M\,\mathbb{P}\left[q(\bar{\boldsymbol{X}}, \boldsymbol{Y}) \geq q(\boldsymbol{X}, \boldsymbol{Y})\,\big|\,\boldsymbol{X}, \boldsymbol{Y}\right]\right\}\right], \quad (14)$$

where $\bar{p}_e^q$ denotes the average error probability under the decoding metric q. For the MMI decoder, we get

$$\bar{p}_e^{\mathrm{mmi}} \dot{=} \mathbb{E}\left[\min\left\{1, M\,\mathbb{P}\left[\frac{I(\hat{P}_{\bar{\boldsymbol{X}}\boldsymbol{Y}})}{I(\hat{P}_{\boldsymbol{X}\boldsymbol{Y}})} \geq 1\,\Big|\,\boldsymbol{X}, \boldsymbol{Y}\right]\right\}\right] \quad (15)$$

$$= \mathbb{E}\left[\min\left\{1, M\,\mathbb{P}\left[\frac{\hat{P}_{\boldsymbol{Y}|\bar{\boldsymbol{X}}}(\boldsymbol{Y}|\bar{\boldsymbol{X}})}{\hat{P}_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X})} \geq 1\,\Big|\,\boldsymbol{X}, \boldsymbol{Y}\right]\right\}\right]. \quad (16)$$

Given that $\min\{1, x\} \leq x^\rho$ for $x \geq 0$ and $\rho \in [0,1]$, we can upper bound the expression for any $\rho \in [0,1]$ as follows

$$\bar{p}_e^{\mathrm{mmi}} \leq \mathbb{E}\left[M^\rho \mathbb{P}\left[\frac{\prod_{i=1}^n \hat{P}_{\boldsymbol{Y}|\bar{\boldsymbol{X}}}(Y_i|\bar{X}_i)}{\prod_{i=1}^n \hat{P}_{\boldsymbol{Y}|\boldsymbol{X}}(Y_i|X_i)} \geq 1\,\Big|\,\boldsymbol{X}, \boldsymbol{Y}\right]^\rho\right] \quad (17)$$

$$\leq \mathbb{E}\left[M^\rho \left(\frac{\mathbb{E}\left[\prod_{i=1}^n \hat{P}_{\boldsymbol{Y}|\bar{\boldsymbol{X}}}(Y_i|\bar{X}_i)\,\big|\,\boldsymbol{Y}\right]}{\prod_{i=1}^n \hat{P}_{\boldsymbol{Y}|\boldsymbol{X}}(Y_i|X_i)}\right)^\rho\right], \quad (18)$$

where in (18) we applied Markov's inequality. Observe that, unlike standard dual-domain derivations, no tilting parameter s is introduced. For a given $\boldsymbol{y}$, [4, Lemma 2.1] allows us to rewrite the expectation in the numerator of (18) as if the distribution were i.i.d. with Q_X, as follows

$$\mathbb{E}\left[\hat{P}_{\boldsymbol{y}|\bar{\boldsymbol{X}}}\left(\boldsymbol{y}|\bar{\boldsymbol{X}}\right)\right] \dot{\leq} \sum_{\bar{\boldsymbol{x}} \in \mathcal{T}^n(Q_X)} \prod_{i=1}^n Q_X(\bar{x}_i)\hat{P}_{\boldsymbol{y}|\bar{\boldsymbol{x}}}(y_i|\bar{x}_i). \quad (19)$$

Recall that $Q_X = \hat{P}_{\bar{\boldsymbol{x}}}$, and note that for $a \in \mathcal{X}$ and $b \in \mathcal{Y}$, we have $\hat{P}_{\bar{\boldsymbol{x}}}(a)\hat{P}_{\boldsymbol{y}|\bar{\boldsymbol{x}}}(b|a) = \hat{P}_{\boldsymbol{y}}(b)\hat{P}_{\bar{\boldsymbol{x}}|\boldsymbol{y}}(a|b)$. Thus, the RHS of (19) can be rewritten as

$$\mathbb{E}\left[\prod_{i=1}^n \hat{P}_{\boldsymbol{y}|\bar{\boldsymbol{X}}}\left(y_i|\bar{X}_i\right)\right]$$

$$\dot{\leq} \hat{P}_{\boldsymbol{y}}(\boldsymbol{y}) \sum_{\bar{\boldsymbol{x}} \in \mathcal{T}^n(Q_X)} \prod_{i=1}^n \hat{P}_{\bar{\boldsymbol{x}}|\boldsymbol{y}}\left(\bar{x}_i|y_i\right) \qquad (20)$$

$$= \hat{P}_{\boldsymbol{y}}(\boldsymbol{y}) \sum_{\substack{\bar{P}_{\bar{X}Y}: \\ \bar{P}_{\bar{X}} = Q_X, \bar{P}_Y = \hat{P}_{\boldsymbol{y}}}} \sum_{\bar{\boldsymbol{x}}:(\bar{\boldsymbol{x}}, \boldsymbol{y}) \in \mathcal{T}^n(\bar{P})} \prod_{i=1}^n \bar{P}_{\bar{X}|Y}(\bar{x}_i|y_i) \quad (21)$$

$$\doteq \hat{P}_{\boldsymbol{y}}(\boldsymbol{y}) \sum_{\substack{\bar{P}_{\bar{X}Y}: \\ \bar{P}_{\bar{X}}=Q_X, \bar{P}_Y=\hat{P}_{\boldsymbol{y}}}} 1 \tag{22}$$

$$\doteq \hat{P}_{\boldsymbol{y}}(\boldsymbol{y}), \tag{23}$$

where $\bar{P}_{\bar{X}|Y}$ in (21) is the conditional distribution of $\bar{X}$ given Y induced by the joint distribution $\bar{P}_{\bar{X}Y}$. The 1 in (22) appears because the product $\prod_{i=1}^{n} \bar{P}_{\bar{X}|Y}(\bar{x}_i|y_i)$ simplifies to $\exp(-nH_{\bar{P}}(\bar{X}|Y))$, while the size of the set $\{\bar{\boldsymbol{x}} : (\bar{\boldsymbol{x}}, \boldsymbol{y}) \in \mathcal{T}^n(\bar{P})\} \doteq \exp(nH_{\bar{P}}(\bar{X}|Y))$, and these terms cancel each other. Substituting this result into (18), we obtain

$$\bar{p}_e^{\text{mmi}} \dot{\leq} \mathbb{E}\left[M^{\rho}\left(\frac{\prod_{i=1}^{n}\hat{P}_{\boldsymbol{Y}}(Y_i)}{\prod_{i=1}^{n}\hat{P}_{\boldsymbol{Y}|\boldsymbol{X}}(Y_i|X_i)}\right)^{\rho}\right] \tag{24}$$

$$= \mathbb{E}\left[M^{\rho}\left(\frac{\prod_{i=1}^{n}\hat{P}_{\boldsymbol{X}}(X_i)}{\prod_{i=1}^{n}\hat{P}_{\boldsymbol{X}|\boldsymbol{Y}}(X_i|Y_i)}\right)^{\rho}\right]. \tag{25}$$

For a given pair $(\boldsymbol{x}, \boldsymbol{y})$, the conditional type can be written as

$$\prod_{i=1}^{n}\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(x_i|y_i) = \sup_U \prod_{i=1}^{n} U(x_i|y_i). \tag{26}$$

To see this, note that for any $U \gg \hat{P}_{\boldsymbol{x}|\boldsymbol{y}}$,

$$\prod_{i=1}^{n}\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(x_i|y_i) = \left(\prod_{i=1}^{n} U(x_i|y_i)\right)\left(\prod_{i=1}^{n}\frac{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(x_i|y_i)}{U(x_i|y_i)}\right). \tag{27}$$

Now, observe that the second term can be expressed as

$$\prod_{i=1}^{n}\frac{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(x_i|y_i)}{U(x_i|y_i)} = \prod_{(a,b)}\left[\frac{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(a|b)}{U(a|b)}\right]^{n\hat{P}_{\boldsymbol{x}\boldsymbol{y}}(a,b)} \tag{28}$$

$$= \exp\left(nD(\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}\|U|\hat{P}_{\boldsymbol{y}})\right), \tag{29}$$

where $D(\cdot \| \cdot | \cdot)$ denotes the conditional relative entropy, as defined in [1, Eq. 2.4]. Hence, we have

$$\prod_{i=1}^{n}\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(x_i|y_i) = \left(\prod_{i=1}^{n} U(x_i|y_i)\right) e^{nD(\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}\|U|\hat{P}_{\boldsymbol{y}})}. \tag{30}$$

Given that the relative entropy is always non-negative and (30) holds for any U, we can take the supremum to obtain

$$\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y}) \geq \sup_U \prod_{i=1}^{n} U(x_i|y_i), \tag{31}$$

with equality achieved when $U = \hat{P}_{\boldsymbol{x}|\boldsymbol{y}}$. Recall that $\hat{P}_{\boldsymbol{x}} = Q_X$ under constant-composition coding. Thus,

$$\bar{p}_e^{\text{mmi}} \dot{\leq} \mathbb{E}\left[M^{\rho}\left(\frac{\prod_{i=1}^{n}Q_X(X_i)}{\sup_U \prod_{i=1}^{n}U(X_i|Y_i)}\right)^{\rho}\right] \tag{32}$$

$$\dot{\leq} \inf_U \mathbb{E}\left[e^{n\rho R}\left(\frac{\prod_{i=1}^{n}Q_X(X_i)}{\prod_{i=1}^{n}U(X_i|Y_i)}\right)^{\rho}\right], \tag{33}$$

where the dot in (33) arises because $M \doteq \exp(nR)$. The upper bound is obtained by moving the minimization outside the expectation. To see this, observe that for any $\bar{\boldsymbol{x}} \in \mathcal{X}^n, \bar{\boldsymbol{y}} \in \mathcal{Y}^n$, and $\bar{U}$, we have

$$\inf_U \left(\frac{Q^n(\bar{\boldsymbol{x}})}{U^n(\bar{\boldsymbol{x}}|\bar{\boldsymbol{y}})}\right)^{\rho} \leq \left(\frac{Q^n(\bar{\boldsymbol{x}})}{\bar{U}^n(\bar{\boldsymbol{x}}|\bar{\boldsymbol{y}})}\right)^{\rho}. \tag{34}$$

The result then follows by taking the expectation with respect to $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$ and subsequently minimizing over $\bar{U}$ on the RHS. Taking advantage of constant-composition coding, we simplify the expectation as follows

$$\mathbb{E}\left[\left(\frac{\prod_{i=1}^{n}Q_X(X_i)}{\prod_{i=1}^{n}U(X_i|Y_i)}\right)^{\rho}\right]$$

$$= \sum_{\boldsymbol{x}\in\mathcal{T}^n(Q_X)}\frac{1}{|\mathcal{T}^n(Q_X)|}\sum_{\boldsymbol{y}}\prod_{i=1}^{n}W(y_i|x_i)\left(\frac{Q_X(x_i)}{U(x_i|y_i)}\right)^{\rho} \tag{35}$$

$$= \sum_{\boldsymbol{x}\in\mathcal{T}^n(Q_X)}\frac{1}{|\mathcal{T}^n(Q_X)|}\prod_{i=1}^{n}\sum_{y}W(y|x_i)\left(\frac{Q_X(x_i)}{U(x_i|y)}\right)^{\rho} \tag{36}$$

$$= \sum_{\boldsymbol{x}\in\mathcal{T}^n(Q_X)}\frac{1}{|\mathcal{T}^n(Q_X)|}\prod_{a\in\mathcal{X}}\left[\sum_{y}W(y|a)\left(\frac{Q_X(a)}{U(a|y)}\right)^{\rho}\right]^{nQ_X(a)} \tag{37}$$

$$= \prod_{x\in\mathcal{X}}\left[\sum_{y}W(y|x)\left(\frac{Q_X(x)}{U(x|y)}\right)^{\rho}\right]^{nQ_X(x)}, \tag{38}$$

where (38) holds because the product term in (37) is identical for all $\boldsymbol{x} \in \mathcal{T}^n(Q_X)$. Accordingly, we have for any $\rho \in [0,1]$,

$$\bar{p}_e^{\text{mmi}} \dot{\leq} \exp\left(-n\left[E_0^{\text{mmi}}(Q_X, \rho) - \rho R\right]\right), \tag{39}$$

where

$$E_0^{\text{mmi}}(Q_X, \rho)$$
$$= \sup_U -\sum_x Q_X(x)\log\sum_y W(y|x)\left(\frac{Q_X(x)}{U(x|y)}\right)^{\rho}. \tag{40}$$

For any non-negative function f defined on $\mathcal{X}$, we have $\mathbb{E}[\log f(X)] = \inf_{r(.)}\log\mathbb{E}\left[\frac{f(X)}{e^{r(X)-\phi_r}}\right]$, where $\phi_r = \mathbb{E}[r(X)]$ [7, Section 2.4.2]. With this approach, the outer expectation in (40) can be moved inside the logarithm, resulting in

$$E_0^{\text{mmi}}(Q_X, \rho)$$
$$= \sup_U \sup_{r(.)} -\log\sum_{(x,y)}Q_X(x)W(y|x)\left(\frac{Q_X(x)}{U(x|y)e^{r(x)-\phi_r}}\right)^{\rho} \tag{41}$$

$$= \sup_{r(.)} -\log\inf_U \sum_{(x,y)}Q_X(x)W(y|x)\left(\frac{Q_X(x)}{U(x|y)e^{r(x)-\phi_r}}\right)^{\rho}. \tag{42}$$

To solve the optimization over U in (42), we move the minimization inside the summation, yielding

$$\inf_U \sum_{(x,y)}Q_X(x)W(y|x)\left(\frac{Q_X(x)}{U(x|y)e^{r(x)-\phi_r}}\right)^{\rho}$$

$$\geq \sum_y \inf_{U(.|y)}\sum_x Q_X(x)W(y|x)\left(\frac{Q_X(x)}{U(x|y)e^{r(x)-\phi_r}}\right)^{\rho}. \tag{43}$$

The expression is convex in U, allowing us to apply [3, Theorem 4.4.1]. This theorem guarantees that, for $x \in \mathcal{X}$ and some λ, the optimal $U^\star(.|y)$ satisfies

$$Q_X(x)W(y|x)\left(\frac{Q_X(x)}{e^{r(x)-\phi_r}}\right)^\rho U^\star(x|y)^{-(1+\rho)} = \lambda, \quad (44)$$

resulting in the optimal solution

$$U^\star(x|y) = \frac{Q_X(x)W(y|x)^{\frac{1}{1+\rho}}e^{\frac{-\rho}{1+\rho}[r(x)-\phi_r]}}{\sum_{\tilde{x}}Q_X(\tilde{x})W(y|\tilde{x})^{\frac{1}{1+\rho}}e^{\frac{-\rho}{1+\rho}[r(\tilde{x})-\phi_r]}}. \quad (45)$$

Note that $U^\star$ is indeed the solution to the optimization problem in (41). Setting $U = U^\star$ in (41) provides a valid lower bound, as $U^\star$ is potentially sub-optimal. However, as demonstrated, $U^\star$ also satisfies the upper bound. Since the lower and upper bounds are identical, this confirms that $U^\star$ is the optimal solution. Thus, by substituting $U^\star$, we obtain

$$\inf_U \sum_{(x,y)} Q_X(x)W(y|x)\left(\frac{Q_X(x)}{U(x|y)e^{r(x)-\phi_r}}\right)^\rho$$
$$= \sum_y \left[\sum_x Q_X(x)W(y|x)^{\frac{1}{1+\rho}}e^{\frac{-\rho}{1+\rho}[r(x)-\phi_r]}\right]^{1+\rho}. \quad (46)$$

By redefining $r(x) := \frac{-\rho}{1+\rho}r(x)$, the optimization over $r(.)$ remains unchanged. Consequently, we can express $E_0^{\text{mmi}}(Q_X, \rho)$ as

$$E_0^{\text{mmi}}(Q_X, \rho)$$
$$= \sup_{r(.)} -\log\sum_y\left[\sum_x Q_X(x)W(y|x)^{\frac{1}{1+\rho}}e^{r(x)-\phi_r}\right]^{1+\rho}, \quad (47)$$

which is equal to $E_0^{\text{ml}}(Q_X, \rho)$, as given in [8, Eq. (53)]. $\quad\square$

IV. JOINT SOURCE CHANNEL CODING

We now consider the transmission of non-equiprobable messages with distribution $P^k(\boldsymbol{v}) = \prod_{i=1}^k P_V(v_i)$, where $\boldsymbol{v} = (v_1, \dots, v_k) \in \mathcal{V}^k$ is the source message, and $\mathcal{V}$ is a finite discrete alphabet. The source message set $\mathcal{V}^k$ is partitioned into N_k disjoint subsets (or classes) $\mathcal{A}_k^{(i)}$, $i = 1, \dots, N_k$, such that $\bigcup_{i=1}^{N_k}\mathcal{A}_k^{(i)} = \mathcal{V}^k$, where N_k can grow sub-exponentially with k. We assume that each class $\mathcal{A}_k^{(i)}$, $i = 1, \dots, N_k$ includes one or more full type classes. For each source message $\boldsymbol{v}$ in $\mathcal{A}_k^{(i)}$, the codewords $\boldsymbol{x}(\boldsymbol{v}) \in \mathcal{X}^n$ are generated independently and uniformly from the type class $\mathcal{T}^n(Q_i)$, where the size of the type class must satisfy $|\mathcal{T}^n(Q_i)| \geq |\mathcal{A}_k^{(i)}|$ for all $i = 1, \dots, N_k$. The work in [9, Theorem 1] derives the MAP random coding error exponents for an arbitrary partition of $\mathcal{V}^k$ and recovers Csiszár's exponent [5]. We next show that the exponent of [9, Theorem 1] can be achieved with the generalized MMI decoder, introduced by Csiszár in [5]. Given the channel output $\boldsymbol{y}$, the decoder selects the source message $\hat{\boldsymbol{v}}$ as follows

$$\hat{\boldsymbol{v}} = \operatorname*{argmax}_{\boldsymbol{v}} I(\hat{P}_{\boldsymbol{x}(\boldsymbol{v})\boldsymbol{y}}) - tH(\hat{P}_{\boldsymbol{v}}), \quad (48)$$

where $t \triangleq \frac{k}{n}$. Similarly to the MMI decoder in channel coding, this decoder does not require knowledge of the source or channel distribution. Following similar steps as in (5)-(11), it can be shown that (48) is equivalent to

$$\hat{\boldsymbol{v}} = \operatorname*{argmax}_{\boldsymbol{v}}\left(\prod_{\ell=1}^k \hat{P}_{\boldsymbol{v}}(v_\ell)\right)\left(\prod_{\ell=1}^n \hat{P}_{\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{v})}(y_\ell|x_{\boldsymbol{v},\ell})\right). \quad (49)$$

Theorem 2. *For a given partition $\mathcal{A}_k^{(i)}$, $i = 1, \dots, N_k$, and associated random-coding distributions Q_i, the average probability of error satisfies*

$$\bar{p}_e^{\text{mmi}} \dot{\leq} \sum_{i=1}^{N_k}\exp\left(-\max_{\rho_i\in[0,1]}\left\{nE_0^{\text{ml}}(Q_i, \rho_i) - E_s^{(i)}(\rho_i, P^k)\right\}\right) \quad (50)$$

where $E_s^{(i)}(\rho, P^k) \triangleq \log\left(\sum_{\boldsymbol{v}\in\mathcal{A}_k^{(i)}}P^k(\boldsymbol{v})^{\frac{1}{1+\rho}}\right)^{1+\rho}$.

Proof. We begin with the extended RCU bound for joint source-channel coding, given in Eq. 26 of [9], and adapt it to the generalized MMI decoder as follows

$$\bar{p}_e^{\text{mmi}} \leq \sum_{i=1}^{N_k}\bar{\varepsilon}_i, \quad (51)$$

where

$$\bar{\varepsilon}_i = \sum_{\boldsymbol{v}\in\mathcal{A}_k^{(i)}}P^k(\boldsymbol{v})\sum_{(\boldsymbol{x},\boldsymbol{y})}P_{\boldsymbol{X}}^{(i)}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})$$
$$\times \min\left\{1, \sum_{j=1}^{N_k}\sum_{\bar{\boldsymbol{v}}\in\mathcal{A}_k^{(j)}}\sum_{\substack{\bar{\boldsymbol{x}}:\hat{P}_{\bar{\boldsymbol{v}}}(\bar{\boldsymbol{v}})\hat{P}_{\boldsymbol{y}|\bar{\boldsymbol{x}}}(\boldsymbol{y}|\bar{\boldsymbol{x}})\\\geq\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x})}}P_{\boldsymbol{X}}^{(j)}(\bar{\boldsymbol{x}})\right\}, \quad (52)$$

and $P_{\boldsymbol{X}}^{(i)}(\boldsymbol{x})$ denotes the constant-composition ensemble distribution, as defined in (3) for Q_i. Next, we use Markov's inequality, once again without applying tilting, to obtain

$$\sum_{\substack{\bar{\boldsymbol{x}}:\hat{P}_{\bar{\boldsymbol{v}}}(\bar{\boldsymbol{v}})\hat{P}_{\boldsymbol{y}|\bar{\boldsymbol{x}}}(\boldsymbol{y}|\bar{\boldsymbol{x}})\\\geq\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x})}}P_{\boldsymbol{X}}^{(j)}(\bar{\boldsymbol{x}}) \leq \sum_{\bar{\boldsymbol{x}}}P_{\boldsymbol{X}}^{(j)}(\bar{\boldsymbol{x}})\left(\frac{\hat{P}_{\bar{\boldsymbol{v}}}(\bar{\boldsymbol{v}})\hat{P}_{\boldsymbol{y}|\bar{\boldsymbol{x}}}(\boldsymbol{y}|\bar{\boldsymbol{x}})}{\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x})}\right). \quad (53)$$

We can again follow similar steps as in (19)-(23), using [4, Lemma 2.1], to obtain

$$\sum_{\bar{\boldsymbol{x}}}P_{\boldsymbol{X}}^{(j)}(\bar{\boldsymbol{x}})\hat{P}_{\boldsymbol{y}|\bar{\boldsymbol{x}}}(\boldsymbol{y}|\bar{\boldsymbol{x}}) \dot{\leq} \hat{P}_{\boldsymbol{y}}(\boldsymbol{y}). \quad (54)$$

Thus, we have

$$\bar{\varepsilon}_i \dot{\leq} \sum_{\boldsymbol{v}\in\mathcal{A}_k^{(i)}}P^k(\boldsymbol{v})\sum_{(\boldsymbol{x},\boldsymbol{y})}P_{\boldsymbol{X}}^{(i)}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})$$
$$\times \min\left\{1, \left(\sum_{j=1}^{N_k}\sum_{\bar{\boldsymbol{v}}\in\mathcal{A}_k^{(j)}}\hat{P}_{\bar{\boldsymbol{v}}}(\bar{\boldsymbol{v}})\right)\frac{\hat{P}_{\boldsymbol{y}}(\boldsymbol{y})}{\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x})}\right\}. \quad (55)$$

Let Λ_i denote the set of all possible types of source sequences in $\mathcal{A}_k^{(i)}$ (i.e., $\Lambda_i \subset \mathcal{P}_k(\mathcal{V})$). Consequently, the summation in the minimum function simplifies to

$$\sum_{j=1}^{N_k} \sum_{\bar{\boldsymbol{v}} \in \mathcal{A}_k^{(j)}} \hat{P}_{\bar{\boldsymbol{v}}}(\bar{\boldsymbol{v}}) = \sum_{j=1}^{N_k} \sum_{\bar{P} \in \Lambda_j} \sum_{\bar{\boldsymbol{v}} \in \mathcal{T}^k(\bar{P})} \hat{P}_{\bar{\boldsymbol{v}}}(\bar{\boldsymbol{v}}) \tag{56}$$

$$= \sum_{j=1}^{N_k} \sum_{\bar{P} \in \Lambda_j} \sum_{\bar{\boldsymbol{v}} \in \mathcal{T}^k(\bar{P})} \exp(-kH(\bar{P})) \tag{57}$$

$$\doteq \sum_{j=1}^{N_k} \sum_{\bar{P} \in \Lambda_j} 1 \tag{58}$$

$$\doteq 1, \tag{59}$$

where the 1 in (58) arises from the fact that $|\mathcal{T}^k(\bar{P})| \doteq \exp(kH(\bar{P}))$, and the final dot-equality holds because the number of possible types in $\mathcal{A}_k^{(j)}$ and N_k both grow sub-exponentially with k. Hence, similarly to (25) we have

$$\bar{\varepsilon}_i = \sum_{\boldsymbol{v} \in \mathcal{A}_k^{(i)}} P^k(\boldsymbol{v}) \sum_{(\boldsymbol{x},\boldsymbol{y})} P_X^{(i)}(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x})$$
$$\times \min\left\{ 1, \frac{\hat{P}_{\boldsymbol{x}}(\boldsymbol{x})}{\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})} \right\}. \tag{60}$$

Observe that $\hat{P}_{\boldsymbol{x}} = Q_i$, and recall the inequality $\min\{1, x\} \leq x^\rho$ for $x \geq 0$ and $\rho \in [0, 1]$. Hence, for any $\rho_i \in [0, 1]$,

$$\bar{\varepsilon}_i \dot{\leq} \sum_{\boldsymbol{v} \in \mathcal{A}_k^{(i)}} P^k(\boldsymbol{v}) \sum_{(\boldsymbol{x} \in \mathcal{T}^n(Q_i),\boldsymbol{y})} \frac{W^n(\boldsymbol{y}|\boldsymbol{x})}{|\mathcal{T}^n(Q_i)|} \left(\frac{Q_i^n(\boldsymbol{x})}{\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})} \right)^{\rho_i}. \tag{61}$$

Recall the equivalent representation of $\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})$ given in (26). Substituting this representation, we can further upper bound $\bar{\varepsilon}_i$ as follows

$$\bar{\varepsilon}_i \dot{\leq} \left(\sum_{\boldsymbol{v} \in \mathcal{A}_k^{(i)}} \frac{P^k(\boldsymbol{v})}{\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})^{\rho_i}} \right)$$
$$\times \left(\inf_U \sum_{(\boldsymbol{x} \in \mathcal{T}^n(Q_i),\boldsymbol{y})} \frac{W^n(\boldsymbol{y}|\boldsymbol{x})}{|\mathcal{T}^n(Q_i)|} \left(\frac{Q_i^n(\boldsymbol{x})}{U^n(\boldsymbol{x}|\boldsymbol{y})} \right)^{\rho_i} \right). \tag{62}$$

Observe that the second term in (62) is identical to the expression in (35), and thus,

$$\bar{\varepsilon}_i \dot{\leq} \left(\sum_{\boldsymbol{v} \in \mathcal{A}_k^{(i)}} \frac{P^k(\boldsymbol{v})}{\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})^{\rho_i}} \right) \exp(-nE_0^{\mathrm{ml}}(Q_i, \rho_i)). \tag{63}$$

It remains to be shown that the first term in (63) is upper bounded by $\exp\left(E_s^{(i)}(\rho_i, P^k) \right)$. To this end, we have

$$\sum_{\boldsymbol{v} \in \mathcal{A}_k^{(i)}} \frac{P^k(\boldsymbol{v})}{\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})^{\rho_i}} = \sum_{\bar{P} \in \Lambda_i} \sum_{\boldsymbol{v} \in \mathcal{T}^k(\bar{P})} \frac{P^k(\boldsymbol{v})}{\bar{P}^k(\boldsymbol{v})^{\rho_i}}. \tag{64}$$

Observe that for $\boldsymbol{v} \in \mathcal{T}^k(\bar{P})$, we have $\frac{P^k(\boldsymbol{v})}{\bar{P}^k(\boldsymbol{v})^{\rho_i}} = \exp\left(k \sum_{b \in \mathcal{V}} \bar{P}(b) \log \frac{P_V(b)}{\bar{P}(b)^{\rho_i}} \right)$, and $|\mathcal{T}^k(\bar{P})| \doteq \exp(kH(\bar{P}))$. Consequently, (64) simplifies to

$$\sum_{\boldsymbol{v} \in \mathcal{A}_k^{(i)}} \frac{P^k(\boldsymbol{v})}{\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})^{\rho_i}} \doteq \sum_{\bar{P} \in \Lambda_i} \exp\left(k \sum_{b \in \mathcal{V}} \bar{P}(b) \log \frac{P_V(b)}{\bar{P}(b)^{1+\rho_i}} \right). \tag{65}$$

The exponential term on the RHS of (65) can be rewritten as

$$\exp\left(k \sum_{b \in \mathcal{V}} \bar{P}(b) \log \frac{P_V(b)}{\bar{P}(b)^{1+\rho_i}} \right)$$
$$= \left[\exp\left(k \sum_{b \in \mathcal{V}} \bar{P}(b) \log \frac{P_V(b)^{\frac{1}{1+\rho_i}}}{\bar{P}(b)} \right) \right]^{1+\rho_i} \tag{66}$$
$$= \left[\sum_{\boldsymbol{v} \in \mathcal{T}^k(\bar{P})} P^k(\boldsymbol{v})^{\frac{1}{1+\rho_i}} \right]^{1+\rho_i}. \tag{67}$$

Substituting this back into (64), we have

$$\sum_{\boldsymbol{v} \in \mathcal{A}_k^{(i)}} \frac{P^k(\boldsymbol{v})}{\hat{P}_{\boldsymbol{v}}(\boldsymbol{v})^{\rho_i}} = \sum_{\bar{P} \in \Lambda_i} \left[\sum_{\boldsymbol{v} \in \mathcal{T}^k(\bar{P})} P^k(\boldsymbol{v})^{\frac{1}{1+\rho_i}} \right]^{1+\rho_i} \tag{68}$$
$$\leq \left[\sum_{\bar{P} \in \Lambda_i} \sum_{\boldsymbol{v} \in \mathcal{T}^k(\bar{P})} P^k(\boldsymbol{v})^{\frac{1}{1+\rho_i}} \right]^{1+\rho_i} \tag{69}$$
$$= \left[\sum_{\boldsymbol{v} \in \mathcal{A}_k^{(i)}} P^k(\boldsymbol{v})^{\frac{1}{1+\rho_i}} \right]^{1+\rho_i}, \tag{70}$$

where in (69), we used $\sum_i a_i^r \leq (\sum_i a_i)^r$ for $r \geq 1$ [3, Problem 4.15.f]. By combining these results and minimizing over ρ_i, we obtain

$$\bar{\varepsilon}_i \dot{\leq} \exp\left(-\max_{\rho_i \in [0,1]} \left\{ nE_0^{\mathrm{ml}}(Q_i, \rho_i) - E_s^{(i)}(\rho_i, P^k) \right\} \right), \tag{71}$$

thus concluding the proof. □

REFERENCES

[1] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.
[2] R. G. Gallager, "Fixed composition arguments and lower bounds to error probability," 1994, available online: http://web.mit.edu/gallager/www/notes/notes5.pdf.
[3] ——, *Information Theory and Reliable Communication*. USA: John Wiley & Sons, Inc., 1968.
[4] G. S. Poltyrev, "Random coding bounds for discrete memoryless channels," *Probl. Peredachi Inf.*, vol. 18, pp. 12–26, 1982.
[5] I. Csiszár, "Joint source-channel error exponent," *Prob. Control Inf. Theory*, vol. 9, no. 1, pp. 315–328, 1980.
[6] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Ensemble-tight error exponents for mismatched decoders," in *2012 50th Allerton Conference on Communication, Control, and Computing*, 2012, pp. 1951–1958.
[7] J. Scarlett, "Reliable communication under mismatched decoding," Ph.D. dissertation, University of Cambridge, 2014.
[8] 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, 2014.
[9] 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, 2014.