A Lagrange–Dual Lower Bound to the Error Exponent Function of the Typical Random Code*

Neri Merhav

The Andrew & Erna Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
E–mail: merhav@ee.technion.ac.il

Abstract

A Lagrange–dual (Gallager–style) lower bound is derived for the error exponent function of the typical random code (TRC) pertaining to the i.i.d. random coding ensemble and mismatched stochastic likelihood decoding. While the original expression, derived from the method of types (the Csiszár–style expression) involves minimization over probability distributions defined on the channel input–output alphabets, the new Lagrange–dual formula involves optimization of five parameters, independently of the alphabet sizes. For both stochastic and deterministic mismatched decoding (including maximum likelihood decoding as a special case), we provide a rather comprehensive discussion on the insight behind the various ingredients of this formula and describe how its behavior varies as the coding rate exhausts the relevant range. Among other things, it is demonstrated that this expression simultaneously generalizes both the expurgated error exponent function (at zero rate) and the classical random coding exponent function at high rates, where it also meets the sphere–packing bound.

Index Terms: error exponent, typical random code, Lagrange duality, mismatched decoder, likelihood decoder.

*This research was supported by the Israel Science Foundation (ISF), grant no. 137/18.

1 Introduction

In view of the articles by Barg and Forney [3], Nazari [12] and Nazari *et al.* [13], in a more recent paper [9], the exact error exponent function of the typical random code (TRC) for a given discrete memoryless channel (DMC) was derived analytically for the ensemble of fixed–composition codes. The error exponent of the TRC was defined as the limit of the negative normalized *expectation of the logarithm* of the error probability, which is different from the classical random coding exponent, defined as the negative normalized *logarithm of the expectation* of the error probability, where both expectations are with respect to (w.r.t.) the randomness of the code. This study of TRC error exponents was motivated by various considerations. The first is that due to Jensen's inequality, it is always greater than or equal to the random coding error exponent, and it is therefore a more optimistic performance metric than the classical random coding exponent, especially at a certain range of low rates. The second consideration is that whenever a certain measure concentration property holds, it is a more relevant figure of merit, because the code is normally assumed to be randomly selected only once, and then it is used repeatedly. Last but not least, it is coherent with the notion of random–like codes [4], which are considered very good codes.

In [9], an exact single–letter expression was derived for the error exponent function of the TRC assuming a general finite alphabet, discrete memoryless channel (DMC) w.r.t. the ensemble of fixed composition codes and a family of stochastic decoders, referred to as generalized likelihood decoders (GLDs), which includes many relevant deterministic decoders (like the maximum likelihood decoder) as special cases. Among other things, it was shown in [9] (similarly as in [3] and [12]), that the error exponent function of the TRC has the following properties: (i) it agrees with the expurgated exponent at rate zero, (ii) it is smaller than the expurgated exponent, but larger than the random coding exponent at a certain range low rates, and (iii) it coincides with the random coding exponent above a certain rate. Other, more recent follow–up papers related to TRC exponents, include time–varying trellis codes [10], codes for colored Gaussian channel [11], joint source–channel coding [1], and large deviations about the TRC exponent [2].

The error exponent formula of [9, Theorem 1] was derived using the method of types, and as such, it was presented in terms of several (nested) optimizations of a certain information–theoretic expression over some joint probability distributions and conditional distributions whose support

depends on the channel input and output alphabets. We henceforth refer to this expression as the *Csiszár–style* formula. The total number of parameters w.r.t. which this expression should be optimized is $|\mathcal{X}|^2 \cdot |\mathcal{Y}| + (|\mathcal{X}|-1) \cdot |\mathcal{Y}| - 1$, where $|\mathcal{X}|$ and $|\mathcal{Y}|$ are the cardinalities of the input alphabet and the output alphabet, respectively. Even in the simplest case of a binary–input/binary–output channel ($|\mathcal{X}| = |\mathcal{Y}| = 2$), this number is already as large as nine, and it grows, of course, extremely rapidly as the alphabet sizes grow. Moreover, out of this number of parameters, $|\mathcal{X}|^2 \cdot |\mathcal{Y}| - 1$ are associated with minimization and the remaining $(|\mathcal{X}| - 1) \cdot |\mathcal{Y}|$ parameters undergo maximization. We make this distinction because if one is interested merely in guaranteed performance, namely, a valid lower bound to the error exponent, there is complete freedom in choosing the latter parameters in an arbitrary manner (rather than maximizing over them), but there is still a necessity to find the global minimum over the former $|\mathcal{X}|^2|\mathcal{Y}| - 1$ parameters, which is still computationally very demanding even for moderate alphabet sizes.

These facts motivate us to derive the Lagrange–dual of the above–mentioned Csiszár–style formula, a.k.a. the *Gallager–style* formula, which is technically, the main result of this paper. For the sake of simplicity, we derive the Lagrange–dual expression for the ensemble of codes drawn from an i.i.d. distribution, in contrast to the ensemble of fixed composition codes, used in [9]. Although the i.i.d. ensemble cannot be better than the fixed composition ensemble [9, Sect. IV.D], we opt to adopt the former for several reasons.

1. The derivation of the Lagrange–dual expression is considerably easier and simpler for the i.i.d. ensemble than for the fixed–composition ensemble, because it is free of the constraints associated with the fixed composition assumption. While the Lagrange–dual form can be derived for the fixed composition too, the cost of eliminating those fixed–composition constraints is in having many additional parameters for optimization.

2. The i.i.d. ensemble is very important on its own right. It has been investigated in much of the earlier, classical work on error exponents [6]. In particular, it was studied [3] for TRC exponents (among other things), albeit only for the special case of the binary symmetric channel.

3. At least at zero–rate and above the critical rate, there is provably no loss of optimality, at least when the random coding distribution is chosen optimally.

While the resulting Lagrange–dual expression is still not trivial, it is nevertheless computationally preferable by far relative to the Csiszár–style expression, as it is associated with optimizations over five parameters, independently of the alphabet sizes. Four of these optimizations are maximizations, and only one is a minimization. Thus, as described above, for a valid lower bound on the TRC exponent, we have full freedom in choosing the former four parameters, and for only one parameter, it is necessary to conduct a minimization.

But the benefit of the Lagrange–dual, Gallager–style lower bound to the TRC exponent is not limited to computational aspects alone. We also provide a rather comprehensive discussion on the insight behind the various ingredients of this formula and describe how its behavior varies as the coding rate exhausts the relevant range. Among other things, it is demonstrated that this expression simultaneously generalizes both the expurgated error exponent function (at zero rate) and the classical random coding exponent function, at high rates, where it also meets the sphere–packing bound.

The outline of the remaining part of this paper is as follows. In Section 2, we establish notation conventions, define the setup, and provide the necessary background from [9]. In Section 3, we present the main result – the Lagrange–dual lower bound to the TRC exponent, and discuss it. Finally, in Section 4, we prove the Lagrange–dual formula.

2 Notation and Background

2.1 Notation

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors and their realizations will be denoted, respectively, by capital letters and the corresponding lower case letters, both in the bold face font. Their alphabets will be superscripted by their dimensions. For example, the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, ($n$ – positive integer) may take a specific vector value $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathcal{X}^n$, the n–th order Cartesian power of $\mathcal{X}$, which is the alphabet of each component of this vector. Sources and channels will be denoted by the letters P, Q and W, subscripted by the names of the relevant random variables/vectors and their conditionings, if applicable, following the standard notation conventions, e.g., Q_X, $P_{Y|X}$, and

so on. When there is no room for ambiguity, these subscripts will be omitted. The expectation operator with respect to (w.r.t.) a probability distribution Q will be denoted by $\boldsymbol{E}_Q\{\cdot\}$. Again, the subscript will be omitted if the underlying probability distribution is clear from the context. For two positive sequences, $\{a_n\}$ and $\{b_n\}$, the notation $a_n \doteq b_n$ will stand for equality in the exponential scale, that is, $\lim_{n\to\infty} \frac{1}{n}\log\frac{a_n}{b_n} = 0$. Similarly, $a_n \dot{\leq} b_n$ means that $\limsup_{n\to\infty} \frac{1}{n}\log\frac{a_n}{b_n} \leq 0$, and so on. The indicator function of an event $\mathcal{E}$ will be denoted by $\mathcal{I}\{E\}$. The notation $[x]_+$ will stand for $\max\{0, x\}$.

The empirical distribution of a sequence $\boldsymbol{x} \in \mathcal{X}^n$, which will be denoted by $\hat{P}_{\boldsymbol{x}}$, is the vector of relative frequencies of each symbol $x \in \mathcal{X}$ in $\boldsymbol{x}$. Similarly, the joint empirical distribution of a sequence pair, $(\boldsymbol{x}.\boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$, will be denoted by $\hat{P}_{\boldsymbol{x}\boldsymbol{y}}$. The type class of a vector $\boldsymbol{x}^n$ with empirical distribution $\hat{P}_{\boldsymbol{x}} = Q_X$ will be denoted by $\mathcal{T}(Q_X)$, and similarly the type class of a pair of vectors $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ with joint empirical distribution $Q_{X\tilde{X}}$ will be denoted by $\mathcal{T}(Q_{X\tilde{X}})$. For a generic distribution, Q_{XY} (or Q, for short, when there is no risk of ambiguity), we use the following notation for information measures: $H_Q(X)$ – for the entropy of X, $H_Q(X,Y)$ – for the joint entropy, $H_Q(X|Y)$ – for the conditional entropy of X given Y, $I_Q(X;Y)$ – for the mutual information, and similar conventions for other information measures and for joint distributions of more than two random variables. We will also use the customary notation for the weighted divergence,

$$D(Q_{Y|X}\|P_{Y|X}|Q_X) = \sum_{x\in\mathcal{X}} Q_X(x) \sum_{y\in\mathcal{Y}} Q_{Y|X}(y|x)\log\frac{Q_{Y|X}(y|x)}{P_{Y|X}(y|x)}. \tag{1}$$

2.2 Background

Consider a DMC, $W = \{W(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$, where $\mathcal{X}$ is a finite input alphabet, $\mathcal{Y}$ is a finite output alphabet, and $W(y|x)$ is the channel input–output single–letter transition probability from x to y. When fed by a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$, the channel responds by producing an output vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathcal{Y}^n$, according to

$$W(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1} W(y_i|x_i). \tag{2}$$

Let $\mathcal{C}_n = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M-1}\} \subseteq \mathcal{X}^n$, $M = e^{nR}$, R being the coding rate in nats per channel use. When the transmitter wishes to convey a message $m \in \{0, 1, \ldots, M-1\}$, it feeds the channel with $\boldsymbol{x}_m$. In [9], we considered the ensemble of fixed composition codes, where each codeword is selected

independently at random under the uniform distribution across a given type class of n–vectors, $\mathcal{T}(Q_X)$.

As in [8] and [9], we adopt an extended version of the so called *likelihood decoder* [14], [15], [17], which is a stochastic decoder that randomly selects the message estimate according to the posterior probability distribution given $\boldsymbol{y}$. The generalized likelihood decoder (GLD) randomly selects the decoded message $\hat{m}$ according to the generalized posterior,

$$P(\hat{m}=m|\boldsymbol{y})=\frac{\exp\{ng(\hat{P}_{\boldsymbol{x}_m\boldsymbol{y}})\}}{\sum_{m'=0}^{M-1}\exp\{ng(\hat{P}_{\boldsymbol{x}_{m'}\boldsymbol{y}})\}}, \tag{3}$$

where the function $g(\cdot)$, henceforth referred to as the *decoding metric*, is a continuous function that maps joint probability distributions over $\mathcal{X}\times\mathcal{Y}$ to real numbers. Thus, considering the function $g(Q)=\beta\cdot\boldsymbol{E}_Q\ln W(Y|X)$ (for a given $\beta>0$), the choice $\beta=1$ corresponds to the ordinary posterior of m given $\boldsymbol{y}$, and the limit $\beta\to\infty$ yields the deterministic maximum a–posteriori (MAP) decoder, which is also the maximum likelihood (ML) decoder in this case. The choice $g(Q_{XY})=\beta\cdot\boldsymbol{E}_Q\ln\tilde{W}(Y|X)$, where $\tilde{W}$ is a possibly different channel, corresponds to a family of stochastic mismatched decoders, which we will adopt throughout this paper. Once again, the limit $\beta\to\infty$ gives a deterministic decoder, in this case, a mismatched decoder. Other interesting choices of g were discussed in [8], [9] as well as in other earlier works.

The probability of error, associated with a given code $\mathcal{C}_n$ and the GLD, is given by

$$P_{\mathrm{e}}(\mathcal{C}_n)=\frac{1}{M}\sum_{m=0}^{M-1}\sum_{m'\neq m}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}W(\boldsymbol{y}|\boldsymbol{x}_m)\cdot\frac{\exp\{ng(\hat{P}_{\boldsymbol{x}_{m'}\boldsymbol{y}})\}}{\sum_{\tilde{m}=0}^{M-1}\exp\{ng(\hat{P}_{\boldsymbol{x}_{\tilde{m}}\boldsymbol{y}})\}}. \tag{4}$$

For the ensemble of rate–R fixed composition codes of type Q_X, we define the TRC error exponent, associated with the decoding metric g, according to

$$E_{\mathrm{trc}}^{g}(R,Q_X)=\lim_{n\to\infty}\left[-\frac{\boldsymbol{E}\ln[P_{\mathrm{e}}(\mathcal{C}_n)]}{n}\right], \tag{5}$$

where the expectations are w.r.t. the randomness of $\mathcal{C}_n$. Note that $E_{\mathrm{trc}}^{g}(R,Q_X)$ is defined in terms of the expectation of the logarithm of the error probability, as opposed to the definition of the ordinary random coding exponent, which is in terms of the logarithm of the expectation of the error probability.

For a given distribution, Q_Y, over the channel output alphabet, let

$$\alpha(R,Q_Y)\stackrel{\Delta}{=}\sup_{\{Q_{\tilde{X}|Y}:\ I_Q(\tilde{X};Y)\leq R,\ Q_{\tilde{X}}=Q_X\}}[g(Q_{\tilde{X}Y})-I_Q(\tilde{X};Y)]+R, \tag{6}$$

and

$$\Gamma(Q_{XX'},R) \stackrel{\Delta}{=} \inf_{Q_{Y|XX'}} \{D(Q_{Y|X}\|W|Q_X) + I_Q(X';Y|X) + [\max\{g(Q_{XY}),\alpha(R,Q_Y)\} - g(Q_{X'Y})]_+\}. \quad (7)$$

The main result of [9] is the following.

Theorem 1 *[9, Theorem 1] Consider the setting described above. Then,*

$$E^g_{\rm trc}(R,Q_X) = \inf_{\{Q_{X'|X}:\ I_Q(X;X')\le 2R,\ Q_{X'}=Q_X\}} \{\Gamma(Q_{XX'},R) + I_Q(X;X') - R\}. \quad (8)$$

As can be seen, the calculation of $\alpha(R,Q_Y)$, which is associated with maximization over $Q_{\tilde{X}|Y}$, involves $|\mathcal{X}|-1$ free parameters for each $y\in\mathcal{Y}$, thus a total of $(|\mathcal{X}|-1)\cdot|\mathcal{Y}$ parameters. The minimizations over $Q_{Y|XX'}$ and $Q_{XX'}$, are equivalent to one minimization over $Q_{XX'Y}$, which has $|\mathcal{X}|^2\cdot|\mathcal{Y}|-1$ free parameters, as described in the Introduction.

3 Main Result and Discussion

3.1 Main Result

We consider the same setting as defined in Section 2, except that the fixed composition ensemble is replaced by the i.i.d. ensemble, where the M codewords are drawn independently, and each one is drawn under the product distribution

$$P(\boldsymbol{x}) = \prod_{i=1}^n P(x_i). \quad (9)$$

The corresponding TRC exponent is defined as in (5), where the expectation is now taken w.r.t. the i.i.d. ensemble defined by P, and it will be denoted by $E^g_{\rm trc}(R,P)$.

Our main result is the following lower bound to the TRC exponent.

Theorem 2 *Consider the setting defined in Section 2, but with the fixed–composition ensemble being replaced by the i.i.d. ensemble defined by P and a GLD with the decoding metric $g(Q) = \beta\boldsymbol{E}_Q\ln\tilde{W}(Y|X)$, for a given $\beta>0$. Then,*

$$E^g_{\rm trc}(R,P) \ge \sup_{0\le\sigma\le\beta}\sup_{0\le\tau\le\beta-\sigma}\inf_{\lambda\ge 0}\sup_{\theta\ge 0}\sup_{\zeta\ge 1+\theta}\left[-\zeta\ln\left\{\sum_{x\in\mathcal{X}}P(x)\left[\sum_{x'\in\mathcal{X}}P(x')\times\right.\right.\right.$$

$$\left(\sum_{y\in\mathcal{Y}} W(y|x) \cdot \frac{\tilde{W}^{\sigma+\tau}(y|x')}{\tilde{W}^{\sigma}(y|x)\left[\sum_{\tilde{x}} P(\tilde{x})\tilde{W}^{1/\lambda}(y|\tilde{x})\right]^{\lambda\tau}} \right)^{1/(1+\theta)} \Bigg]^{(1+\theta)/\zeta} \Bigg\} - (\zeta+\theta-\lambda\tau)R \Big] \,. \tag{10}$$

3.2 Discussion

First, we note that Theorem 2 provides a lower bound to the TRC exponent, unlike Theorem 1 of [9] that claims the exact TRC exponent. The reason is that, in contrast to [9], here, for the i.i.d. ensemble, we have not proved a matching upper bound, because our emphasis in this work is on a dual expression for the guaranteed performance of the typical random code. Also, in following paragraphs of our discussion, we will discuss several specific choices of the parameters σ, τ, θ, and ζ, rather than maximizing upon them, and so, the resulting expression can only be claimed to be a lower bound anyway. Having said that, we will see in the sequel, that at least in certain situations, the resulting quantities will turn out to be tight, as they will meet well known converse bounds. For the matched case, they will also be coherent with results derived in earlier works.

While the lower bound in Theorem 2 seems to be considerably complicated, some useful insights can nevertheless be gained using a few observations.[1] In particular, first observe that the inner–most sum over y can be rewritten slightly differently as follows:

$$\sum_{y} W(y|x) \cdot \left[\frac{\tilde{W}(y|x')}{\tilde{W}(y|x)}\right]^{\sigma} \cdot \left[\frac{\tilde{W}(y|x')}{\left\{\sum_{\tilde{x}} P(\tilde{x})\tilde{W}^{1/\lambda}(y|\tilde{x})\right\}^{\lambda}}\right]^{\tau} . \tag{11}$$

For simplicity, consider the limit $\beta \to \infty$, where the GLD becomes the deterministic mismatched decoder $\hat{m} = \arg\max_m \tilde{W}(\boldsymbol{y}|\boldsymbol{x}_m)$, and then the minimizations over σ and τ both become over all positive reals. Let us think of the error event as the *disjoint* union of the events

$$\left\{\tilde{W}(\boldsymbol{y}|\boldsymbol{x}_{m'}) > \max\left\{\tilde{W}(\boldsymbol{y}|\boldsymbol{x}_m), \max_{\tilde{m}\neq(m,m')} \tilde{W}(\boldsymbol{y}|\boldsymbol{x}_{\tilde{m}})\right\}\right\}, \qquad m' = 1, 2, \ldots, m-1, m+1, \ldots, M$$

where $\boldsymbol{x}_m$ is the correct codeword. In the above expression of the TRC exponent, x represents the correct codeword $\boldsymbol{x}_m$, x' stands for $\boldsymbol{x}_{m'}$, and $\tilde{x}$ designates the codeword $\boldsymbol{x}_{\tilde{m}}$ with the highest score among all competing wrong codewords other than $\boldsymbol{x}_{m'}$. The summation over y in (11) can be

[1] It should be noted that there is a certain similarity to the dual expression derived in [1] for Slepian–Wolf binning, but there are also quite a few differences.

thought of as a single–letter version of the Chernoff bound for the probability of the above event, which can be rewritten as

$$\left\{\tilde{W}(\boldsymbol{y}|\boldsymbol{x}_{m'}) > \tilde{W}(\boldsymbol{y}|\boldsymbol{x}_m)\right\} \bigcap \left\{\tilde{W}(\boldsymbol{y}|\boldsymbol{x}_{m'}) > \max_{\tilde{m}\neq(m,m')} \tilde{W}(\boldsymbol{y}|\boldsymbol{x}_{\tilde{m}})\right\}.$$

Low coding rates are characterized by the regime where pairwise error events dominate the error probability. These involve merely the competition between $\boldsymbol{x}_m$ and $\boldsymbol{x}_{m'}$, just like in the simple union bound. In this case, the TRC exponent is achieved for $\tau = 0$, and eq. (11) has the meaning of the expectation (w.r.t. $W(y|x)$) of $[\tilde{W}(y|x')/\tilde{W}(y|x)]^{\sigma}$, as the effect of the event

$$\left\{\tilde{W}(\boldsymbol{y}|\boldsymbol{x}_{m'}) > \max_{\tilde{m}\neq(m,m')} \tilde{W}(\boldsymbol{y}|\boldsymbol{x}_{\tilde{m}})\right\}$$

is negligible and hence ignored. In particular, for matched ML decoding, where $\tilde{W} = W$, the choice $\sigma = \frac{1}{2}$ corresponds to the appearance of the Bhattacharyya distance in this expression. As R grows, pairwise error events gradually cease to dominate the error probability, and the decoded codeword, symbolized by x', has to "compete", not only with the correct codeword, but also with all other codewords at the same time. Indeed, the factor

$$\left[\sum_{\tilde{x}} P(\tilde{x})\tilde{W}^{1/\lambda}(y|\tilde{x})\right]^{\lambda},$$

at the denominator of (11), represents the typical overall collective contribution of all other competing codewords, except the correct one. More precisely, it stands for the typical value of the highest likelihood score among all other wrong codewords, which are drawn randomly and independently of the channel output y. As R grows beyond a certain point, more and more weight is given to this term at the expense of the factor pertaining to the correct codeword x. This means that τ ceases to be equal to zero and it becomes positive. Another effect of increasing R is via the choice of the parameter λ. For $\lambda \to 0$, this factor tends to $\max_{\tilde{x}} \tilde{W}(y|\tilde{x})$, which means that at extremely high coding rates, there is enough probability that one of the wrong randomly generated codewords is composed of the "most likely" input letter for each coordinate of $\boldsymbol{y}$. For $\lambda = 1$, it is equal to $\sum_x P(x)\tilde{W}(y|x)$, which, for the matched case ($\tilde{W} = W$), corresponds to $R = I(X;Y)$, the mutual information induced by P and W. Finally, in the other extreme, $\lambda \to \infty$ yields $\exp\{\sum_x P(x) \ln \tilde{W}(y|x)\}$, which is the typical score of a single randomly chosen codeword, namely, zero rate.

From this point onward, until the last paragraph of this section, our discussion focuses on the matched case, namely, $\tilde{W} = W$ and $\beta \to \infty$. In this case, it is interesting to observe that the above derived expression of the TRC exponent is simultaneously a generalized form of both the random coding exponent function and the form of the expurgated function at low rates. To see this, consider the following two cases.

1. For a given low rate R, let $\varrho \geq 1$ be the achiever of

$$\sup_{\rho \geq 1}[E_{\mathrm{x}}(\rho) - 2\rho R], \tag{12}$$

where

$$E_{\mathrm{x}}(\rho) = \rho \ln \left(\sum_{x,x'} P(x)P(x') \left[\sum_{y} \sqrt{W(y|x)W(y|x')} \right]^{1/\rho} \right), \tag{13}$$

and where by "low rate", we mean $R \leq \frac{R_{\mathrm{x}}}{2} \stackrel{\Delta}{=} \frac{\dot{E}_{\mathrm{x}}(1)}{2}$, $\dot{E}_{\mathrm{x}}(\rho)$ being the derivative of $E_{\mathrm{x}}(\rho)$. Now, let $\sigma = \frac{1}{2}$, $\tau = 0$ (so λ is immaterial), $\zeta = \varrho$ and $\theta = \varrho - 1$. The matched TRC exponent is then lower bounded by

$$\begin{aligned} E^{g}_{\mathrm{trc}}(R,P) &\geq -\varrho \ln \left(\sum_{x,x'} P(x)P(x') \left[\sum_{y} \sqrt{W(y|x)W(y|x')} \right]^{1/\varrho} \right) - (2\varrho - 1)R \\ &= E_{\mathrm{x}}(\varrho) - (2\varrho - 1)R \\ &= \sup_{\rho \geq 1}\{E_{\mathrm{x}}(\rho) - (2\rho - 1)R\} \\ &= E_{\mathrm{ex}}(2R, P) + R, \end{aligned} \tag{14}$$

where $E_{\mathrm{ex}}(\cdot, P)$ is the expurgated exponent function [6], [16]. This is in agreement with the results in [3]. In particular, for $R = 0$, $E_{\mathrm{trc}}(0, P) = E_{\mathrm{ex}}(0, P)$ is achieved for $\varrho \to \infty$, which, for the optimal P, is also the optimal achievable zero–rate error exponent [16, Sect. 3.7].

2. For a given high rate R, let now $\varrho \in (0, 1]$, be the achiever of

$$E_{\mathrm{r}}(R) = \sup_{\rho \geq 0}[E_0(\rho) - \rho R] = \sup_{0 \leq \rho \leq 1}[E_0(\rho) - \rho R], \tag{15}$$

where we recall that

$$E_0(\rho) = -\ln \left(\sum_{y} \left[\sum_{x} P(x) W^{1/(1+\rho)}(y|x) \right]^{1+\rho} \right)$$

$$= \quad -\ln\left[\sum_y \exp\left\{(1+\rho)A(y,1+\rho)\right\}\right] \tag{16}$$

with the function $A(\cdot,\cdot)$ being defined as

$$A(y,r) \stackrel{\Delta}{=} \ln\left[\sum_x P(x)W^{1/r}(y|x)\right], \qquad r>0, \tag{17}$$

and where by "high rate", we mean $R \geq \dot{E}_0(1)$, $\dot{E}_0(\cdot)$ being the derivative of $E_0(\cdot)$. This means that $R = \dot{E}_0(\varrho)$. Now, given ϱ, let us choose the parameters as follows: $\sigma = \frac{\rho}{1+\varrho}$, $\tau = \frac{1-\varrho}{1+\varrho}$, $\zeta = 1$, and $\theta = 0$. We then have

$$E_{\text{trc}}(R,P) \geq \inf_{\lambda \geq 0}\left\{E_1(\varrho,\lambda) - \left[1 - \frac{\lambda(1-\varrho)}{1+\varrho}\right]R\right\}, \tag{18}$$

where

$$\begin{aligned} E_1(\varrho,\lambda) &= -\ln\left\{\sum_y \frac{\left[\sum_x P(x)W^{1/(1+\varrho)}(y|x)\right]^2}{\left[\sum_x P(x)W^{1/\lambda}(y|x)\right]^{\lambda(1-\varrho)/(1+\varrho)}}\right\} \\ &= -\ln\left[\sum_y \exp\left\{2A(y,1+\varrho) - \frac{\lambda(1-\varrho)}{1+\varrho}A(y,\lambda)\right\}\right]. \end{aligned} \tag{19}$$

To find the achiever λ of the r.h.s. of eq. (18), we equate the derivative of the objective to zero, i.e.,

$$\begin{aligned} 0 &= \frac{\partial E_1(\varrho,\lambda)}{\partial\lambda} + \frac{(1-\varrho)R}{1+\varrho} \\ &\equiv \frac{\partial E_1(\varrho,\lambda)}{\partial\lambda} + \frac{1-\varrho}{1+\varrho}\cdot\dot{E}_0(\varrho) \\ &\equiv \frac{1-\varrho}{1+\varrho}\cdot\frac{\sum_y[A(y,\lambda)+\lambda A'(y,\lambda)]\exp\{2A(y,1+\varrho)-\lambda(1-\varrho)A(y,\lambda)/(1+\varrho)\}}{\sum_y \exp\{2A(y,1+\varrho)-\lambda(1-\varrho)A(y,\lambda)/(1+\varrho)\}} - \\ &\quad \frac{1-\varrho}{1+\varrho}\cdot\frac{\sum_y[A(y,1+\varrho)+(1+\varrho)A'(y,1+\varrho)]\exp\{(1+\varrho)A(y,1+\varrho)\}}{\sum_y \exp\{(1+\varrho)A(y,1+\varrho)\}}, \end{aligned} \tag{20}$$

where $A'(y,\lambda)$ is the derivative of $A(y,\lambda)$ w.r.t. λ. It is now easy to see that $\lambda = 1+\varrho$ trivially solves this equation, and so, under the assumption (to be discussed in the next paragraph) that this solution provides the global minimum of the r.h.s. of (18), we have

$$\begin{aligned} E^g_{\text{trc}}(R,P) &\geq \inf_{\lambda\geq 0}\left\{E_1(\varrho,\lambda) - \left[1-\frac{\lambda(1-\varrho)}{1+\varrho}\right]R\right\} \\ &= E_1(\varrho,1+\varrho) - \left[1-\frac{(1+\varrho)(1-\varrho)}{1+\varrho}\right]R \end{aligned}$$

$$
\begin{aligned}
&= E_0(\varrho) - \varrho R \\
&= \sup_{\rho \geq 0} [E_0(\rho) - \rho R] \\
&= E_{\mathrm{r}}(R, P) = E_{\mathrm{sp}}(R, P). \qquad (21)
\end{aligned}
$$

Obviously, the inequality must be achieved with equality, at least for the optimal P, since it coincides with the sphere–packing upper bound to the error exponent.

Referring to the above assumption that $\lambda = 1 + \varrho$ is the minimizer of $E_1(\varrho, \lambda)$ w.r.t. λ, our observations are as follows. A sufficient (though not necessary) condition for this to be the case is that $E_1(\varrho, \lambda)$ would be convex in λ for fixed ϱ. Consider the important special case where $A(y, \lambda)$ is independent[2] of y for all λ. In this case, it is easy to prove the convexity of $E_1(\varrho, \cdot)$ as follows. Abbreviating the notation $A(y, \lambda)$ as $A(\lambda)$, we have

$$
\begin{aligned}
E_1(\varrho, \lambda) &= -\ln \left[\sum_y \exp \left\{ 2A\left(1+\varrho\right) - \frac{\lambda(1-\varrho)}{1+\varrho} A\left(\lambda\right) \right\} \right] \\
&= -\ln \left[|\mathcal{Y}| \cdot \exp \left\{ 2A\left(1+\varrho\right) - \frac{\lambda(1-\varrho)}{1+\varrho} A\left(\lambda\right) \right\} \right] \\
&= \frac{1-\varrho}{1+\varrho} \cdot \lambda A\left(\lambda\right) - \ln |\mathcal{Y}| - \ln \left[2A\left(1+\varrho\right) \right], \qquad (22)
\end{aligned}
$$

which is convex in λ for fixed ϱ because the function $\lambda \cdot A(\lambda)$ is such, as will be shown in Section 4 (see, in particular, eq. (32) and the following text). Beyond the class of channels with the above described symmetry property, a numerical study indicates that $E_1(\varrho, \cdot)$ remains convex for many other combinations of P and W, but not in general. For example, when W is the binary z–channel, this is not always the case.

To summarize, there are basically three ranges with different kinds of behavior of the TRC exponent. Denoting $R_{\mathrm{c}1} = \frac{\dot{E}_{\mathrm{x}}(1)}{2}$ and $R_{\mathrm{c}2} = \dot{E}_0(1)$, we have the following explicit lower bounds to the TRC exponent, which are coherent with the findings of [3], that were derived for the special case of the binary symmetric channel and codes drawn by fair coin tossing.

1. *Low rates.* For $R \leq R_{\mathrm{c}1}$, the graph of the TRC exponent function is a convex curve, with an initial slope of $-\infty$ and final slope of -1. Here, $\sigma = \frac{1}{2}$, $\tau = 0$, and $\zeta = 1 + \theta$ decreases from

[2]This is the case, for example, if P is the uniform distribution and the columns of the matrix $\{w_{ij} = W(j|i)\}$ are permutations of each other. Even more specifically, this happens when $\mathcal{X} = \mathcal{Y}$ is a group and $W(y|x) = W(y \ominus x)$, where $\ominus$ is the difference operation w.r.t. this group, namely, when W is a modulo–additive channel. More generally, this condition continues to hold as long as all columns of W are obtained from one column by permuting only components (of that column) pertaining to channel input symbols for which P assigns the same probability.

∞ to 1 and

$$E^g_{\rm trc}(R,P) \geq E_{\rm ex}(2R,P) + R.$$

2. *Moderate rates.* For $R_{\rm c1} \leq R \leq R_{\rm c2}$, the TRC exponent is an affine function of R with slope -1 (i.e., the graph is a straight line). Here, σ and τ are as before, but $\theta = 0$ and $\zeta = 1$. In this case,

$$E^g_{\rm trc}(R,P) \geq E_0(1) - R.$$

3. *High rates.* For $R \geq R_{\rm c2}$, the graph of the TRC exponent function is again a convex curve, with an initial slope of -1 and final slope of 0. Here, every rate corresponds to a value of ϱ that decreases from 1 to 0, and the parameters are: $\sigma = \frac{\varrho}{1+\varrho}$, $\tau = \frac{1-\varrho}{1+\varrho}$, $\lambda = 1+\varrho$, $\theta = 0$ and $\zeta = 1$, and then

$$E^g_{\rm trc}(R,P) \geq E_{\rm sp}(R,P).$$

It should be pointed out that since the parameters σ, τ, ζ and θ, were chosen here in a specific (seemingly, arbitrary) manner, and not as a result of the maximizations, the resulting expressions are merely lower bounds to the TRC exponent, and there is no guarantee that they are the exact quantities, at all rates, except the cases of $R = 0$ and high rates, above the critical rate, where if P is chosen optimally, these figures meet the well known zero–rate bound and the sphere–packing bound, respectively. However, at intermediate rates, where the exact reliability function is no fully known, it is not clear, for example, that the choice $\tau = 0$ continues to be optimal for all rates up to $R_{\rm c2}$. We could not rule out the theoretical possibility that the passage from $\tau = 0$ to $\tau > 0$, which stands for the point at which pairwise error events cease to dominate the error probability, might occur at a rate strictly lower than $R_{\rm c2}$. On the other hand, numerical studies that we have conducted so far, did not reveal examples where this in fact happens. We therefore conjecture that the optimal value of τ is zero for all $R \leq R_{\rm c2}$.

Our final remark concerns the dependence of the TRC exponent bound on β. It is known [5], [7] that the error probability of the matched likelihood decoder ($\tilde{W} = W$, $\beta = 1$) cannot be larger than twice the error probability of the ML decoder, and therefore, in this case, the optimal error exponent is achieved for every $\beta \geq 1$. It seems to be less obvious, however, that the optimal error exponent of the ML decoder is achieved even if one starts sweeping β from values less than 1. Indeed, as mentioned earlier, when $\beta \to \infty$, the maximization over both σ and τ take place on the

entire positive real line. Let σ^* and τ^* denote the maximizers for $\beta \to \infty$. If σ^* and τ^* are both finite, these maximizers will be achieved as well whenever $\beta \geq \beta_0 = \sigma^* + \tau^*$. Thus, for low rates, if $\sigma^* = \frac{1}{2}$ and $\tau^* = 0$, then $\beta_0 = \frac{1}{2}$ is the critical value of β beyond which the error exponent ceases to improve and remains fixed. For high rates, if $\sigma^* = \frac{\varrho}{1+\varrho}$ and $\tau^* = \frac{1-\varrho}{1+\varrho}$, then $\beta_0 = \frac{1}{1+\varrho}$, so here $\beta_0 \to 1$ only when $R \to I(X;Y)$. Even less obvious is the similar behavior for the mismatched likelihood decoder. The first non–trivial fact about stochastic mismatched decoding is that the error exponent must be a monotonically non–decreasing function of β. The second point is that here too, for the same reasons, if the achievers σ^* and τ^* are finite, the resulting error exponent would cease to depend on β for all $\beta \geq \beta_0 = \sigma^* + \tau^*$. However, here we do not claim that $\beta_0 \leq 1$ in general.

4 Proof

Before passing to the Lagrange–dual, we first need to modify the Csiszár–style expression of the TRC exponent, in Theorem 1 above, in order to account for the fact that we are replacing the fixed–composition ensemble of [9] to the i.i.d. ensemble under P, as described in Section 2. There are only a few modifications. The first is that the constraints $Q_{X'} = Q_X$ (in eq. (6)) and $Q_{\tilde{X}} = Q_X$ (in (8)) are now removed since the types of the randomly chosen codewords may fluctuate around P. The second modification is that the mutual information term, $I_Q(X;X')$, in the objective of (8), and $I_Q(\tilde{X};Y)$, in both the constraint and the objective of (6), are replaced by $J_Q(X;X') + D(Q_X\|P)$ and $J_Q(\tilde{X};Y)$, respectively, where

$$J_Q(X;X') = I_Q(X;X') + D(Q_{X'}\|P) = \boldsymbol{E}_Q \ln \frac{Q_{X'|X}(X'|X)}{P(X')}, \tag{23}$$

and

$$J_Q(\tilde{X};Y) = I_Q(\tilde{X};Y) + D(Q_{\tilde{X}}\|P) = \boldsymbol{E}_Q \ln \frac{Q_{\tilde{X}|Y}(X|Y)}{P(\tilde{X})}. \tag{24}$$

As for the mutual information term in the constraint of (8), the situation is slightly more involved. Referring to the proof of [9, Theorem 1], we have to analyze once again the moments of the type class enumerators,

$$N(Q_{XX'}) = \sum_{m=0}^{M-1} \sum_{m' \neq m} \mathcal{I}\{(\boldsymbol{x}_m, \boldsymbol{x}_{m'}) \in \mathcal{T}(Q_{XX'})\}. \tag{25}$$

Following [9], let us also denote by $N(Q_{XX'}|\boldsymbol{x}_m)$ the number of codewords $\{\boldsymbol{x}_{m'},\ m' \neq m\}$ such that $(\boldsymbol{x}_m, \boldsymbol{x}_{m'}) \in \mathcal{T}(Q_{XX'})$ (i.e., the same definition as $N(Q_{XX'})$ but without the summation over m). Similarly as in [9, eq. (36)], for given $\rho \geq s \geq 1$,

$$\begin{aligned}
\boldsymbol{E}\{N^{1/\rho}(Q_{XX'})\} &= \boldsymbol{E}\left\{\left[\sum_{m=0}^{M-1} N(Q_{XX'}|\boldsymbol{X}_m)\cdot\mathcal{I}\{\boldsymbol{X}_m \in \mathcal{T}(Q_X)\}\right]^{1/\rho}\right\} \\
&= \boldsymbol{E}\left\{\left(\left[\sum_{m=0}^{M-1} N(Q_{XX'}|\boldsymbol{X}_m)\cdot\mathcal{I}\{\boldsymbol{X}_m \in \mathcal{T}(Q_X)\}\right]^{1/s}\right)^{s/\rho}\right\} \\
&\leq \boldsymbol{E}\left\{\left(\sum_{m=0}^{M-1} N^{1/s}(Q_{XX'}|\boldsymbol{X}_m)\cdot\mathcal{I}\{\boldsymbol{X}_m \in \mathcal{T}(Q_X)\}\right)^{s/\rho}\right\} \\
&\leq \left(\boldsymbol{E}\sum_{m=0}^{M-1} N^{1/s}(Q_{XX'}|\boldsymbol{X}_m)\cdot\mathcal{I}\{\boldsymbol{X}_m \in \mathcal{T}(Q_X)\}\right)^{s/\rho} \\
&= e^{nRs/\rho}\left(\boldsymbol{E}\left\{N^{1/s}(Q_{XX'}|\boldsymbol{X}_m)\cdot\mathcal{I}\{\boldsymbol{X}_m \in \mathcal{T}(Q_X)\}\right\}\right)^{s/\rho} \\
&= e^{nRs/\rho}\left(\boldsymbol{E}\left\{N^{1/s}(Q_{XX'}|\boldsymbol{X}_m)\right\}\cdot P[\mathcal{T}(Q_X)]\right)^{s/\rho} \\
&\doteq e^{n[R-D(Q_X\|P)]s/\rho}\left(\boldsymbol{E}\left\{N^{1/s}(Q_{XX'}|\boldsymbol{X}_m)\right\}\right)^{s/\rho} \\
&\doteq e^{n[R-D(Q_X\|P)]s/\rho}\cdot\left\{\begin{array}{ll} e^{n[R-J_Q(X;X')]} & R > J_Q(X;X') \\ e^{n[R-J_Q(X;X')]s/\rho} & R < J_Q(X;X')\end{array}\right. \\
&= \left\{\begin{array}{ll} e^{n\{[R-D(Q_X\|P)]s/\rho+[R-J_Q(X;X')]/\rho\}} & R > J_Q(X;X') \\ e^{n[2R-D(Q_X\|P)-J_Q(X;X')]s/\rho} & R < J_Q(X;X')\end{array}\right.
\end{aligned} \tag{26}$$

and after minimization over $s \in [1, \rho]$:

$$\boldsymbol{E}\{N^{1/\rho}(Q_{XX'})\} \doteq \left\{\begin{array}{ll}
e^{n[2R-D(Q_X\|P)-J_Q(X;X')]/\rho} & R > J_Q(X;X'),\ R > D(Q_X\|P) \\
e^{n[R-D(Q_X\|P)+(R-J_Q(X;X')/\rho]} & R > J_Q(X;X'),\ R < D(Q_X\|P) \\
e^{n[2R-D(Q_X\|P)-J_Q(X;X')]/\rho} & R < J_Q(X;X'),\ 2R > D(Q_X\|P)+J_Q(X;X') \\
e^{n[2R-D(Q_X\|P)-J_Q(X;X')]} & R < J_Q(X;X'),\ 2R < D(Q_X\|P)+J_Q(X;X')
\end{array}\right. \tag{27}$$

Now,

$$\begin{aligned}
&\lim_{\rho\to\infty}\left[\boldsymbol{E}\{N^{1/\rho}(Q_{XX'})\}\right]^{\rho} \\
&\leq \left\{\begin{array}{ll}
e^{n[2R-D(Q_X\|P)-J_Q(X;X')]} & R > J_Q(X;X'),\ R > D(Q_X\|P) \\
0 & R > J_Q(X;X'),\ R < D(Q_X\|P) \\
e^{n[2R-D(Q_X\|P)-J_Q(X;X')]} & R < J_Q(X;X'),\ 2R > D(Q_X\|P)+J_Q(X;X') \\
0 & R < J_Q(X;X'),\ 2R < D(Q_X\|P)+J_Q(X;X')
\end{array}\right.
\end{aligned}$$

$$= \begin{cases} e^{n[2R-D(Q_X\|P)-J_Q(X;X')]} & 2R > F_Q(X,X') \\ 0 & 2R < F_Q(X,X') \end{cases} \tag{28}$$

where

$$F_Q(X,X') = D(Q_X\|P) + \max\{D(Q_X\|P), J_Q(X;X')\}. \tag{29}$$

Using these facts the same way as in [9], we find that the TRC exponent for the i.i.d. ensemble and $g(Q) = \beta \boldsymbol{E}_Q \ln \tilde{W}(Y|X)$, is as follows. We first ref–define $\alpha(R, Q_Y)$ as

$$\alpha(R,Q_Y) \stackrel{\Delta}{=} \sup_{\{Q_{\tilde{X}|Y}:\ J_Q(\tilde{X};Y)\le R\}} [\beta \boldsymbol{E}_Q \ln \tilde{W}(Y|\tilde{X}) - J_Q(\tilde{X};Y)] + R, \tag{30}$$

$\Gamma(Q_{XX'}, R)$ is defined as in (7) (but with $g(Q) = \beta \boldsymbol{E}_Q \ln \tilde{W}(Y|X)$), and finally,

$$E^g_{\rm trc}(R,P) \ge \inf_{\{Q_{X'|X}:\ F_Q(X;X')\le 2R\}} \{\Gamma(Q_{XX'},R) + J_Q(X;X') + D(Q_X\|P) - R\}. \tag{31}$$

We are now ready to move on to the main part of the proof, which is the derivation of the Lagrange–dual form of this lower bound to $E^g_{\rm trc}(R,P)$.

Throughout this proof we will make a frequent use of the minimax theorem, based on convexity–concavity arguments. We will also use repeatedly the fact that for a given function $f : \mathcal{X} \to \mathbb{R}$ that does not depend on Q,

$$\min_Q [D(Q\|P) + \boldsymbol{E}_Q\{f(X)\}] = -\ln \boldsymbol{E}_P\left\{e^{-f(X)}\right\},$$

which can easily be verified either by carrying out the minimization using standard methods, or by writing the the objective on the l.h.s. as

$$D(Q\|P') - \ln \boldsymbol{E}_P\left\{e^{-f(X)}\right\},$$

with $P'(x) \propto P(x)e^{-f(x)}$, which is obviously minimized by $Q = P'$.

We begin from $\alpha(R, Q_Y)$, which is the inner–most optimization.

$$\begin{aligned} \alpha(R,Q_Y) &= \sup_{\{Q_{\tilde{X}|Y}:\ J_Q(\tilde{X};Y)\le R\}} [\beta \boldsymbol{E}_Q \ln \tilde{W}(Y|\tilde{X}) - J_Q(\tilde{X};Y)] + R \\ &= \sup_{Q_{\tilde{X}|Y}} \inf_{\lambda>0} \left[\sum_y Q_Y(y) \sum_x Q_{\tilde{X}|Y}(x|y) \ln \tilde{W}^\beta(y|x) + \right. \\ &\qquad \left. \lambda\left(R - \sum_x Q_{\tilde{X}|Y}(x|y) \ln \frac{Q_{\tilde{X}|Y}(x|y)}{P(x)}\right)\right] \end{aligned}$$

$$
\begin{aligned}
&= \inf_{\lambda>0} \sup_{Q_{\tilde{X}|Y}} \lambda \cdot \left[\sum_y Q_Y(y) \sum_x Q_{\tilde{X}|Y}(x|y) \ln \frac{P(x)\tilde{W}^{\beta/\lambda}(y|x)}{Q_{\tilde{X}|Y}(x|y)} + R \right] \\
&= \inf_{\lambda>0} \lambda \cdot \left[\sum_y Q_Y(y) \ln \left(\sum_x P(x)\tilde{W}^{\beta/\lambda}(y|x) \right) + R \right] \\
&= \inf_{\lambda>0} \lambda \cdot \left[\sum_y Q_Y(y) A\left(y, \frac{\lambda}{\beta}\right) + R \right], \qquad (32)
\end{aligned}
$$

where the function A has been defined in (17). Observe that $\lambda \cdot A(y, \lambda/\beta)$ is always a convex function, since it was obtained as the supremum (over $\{Q_{\tilde{X}|Y}\}$) of affine functions in λ. Now,

$$
\begin{aligned}
E_{\rm trc}^g(R,P) \;\geq\; & \inf_{\{Q_{XX'Y}:\ F_Q(X,X')\leq 2R\}} \{D(Q_{Y|X}\|W|Q_X) + \\
& I_Q(X';Y|X) + J_Q(X;X') + D(Q_X\|P) + \\
& [\max\{\beta \boldsymbol{E}_Q \ln \tilde{W}(Y|X), \alpha(R,Q_Y)\} - \beta \boldsymbol{E}_Q \ln \tilde{W}(Y|X')]_+\} - R \\
= \; & \inf_{\{Q_{XX'Y}:\ F_Q(X,X')\leq 2R\}} \Big\{ - \boldsymbol{E}_Q \ln W(Y|X) - H_Q(Y|X,X') + \\
& J_Q(X;X') + D(Q_X\|P) + \\
& [\max\{\beta \boldsymbol{E}_Q \ln \tilde{W}(Y|X), \alpha(R,Q_Y)\} - \beta \boldsymbol{E}_Q \ln \tilde{W}(Y|X')]_+ \Big\} - R \\
= \; & \inf_{\{Q_{XX'Y}:\ F_Q(X,X')\leq 2R\}} \sup_{0\leq s\leq 1} \sup_{0\leq t\leq 1} \inf_{\lambda>0} \Big\{ J_Q(X;X') + D(Q_X\|P) + \\
& \sum_{x,x'} Q_{XX'}(x,x') \sum_y Q_{Y|XX'}(y|x,x') \left[\ln \frac{Q_{Y|XX'}(y|xx')}{W(y|x)} + \right. \\
& \left. s\left(t \ln \tilde{W}^{\beta}(y|x) + (1-t)\lambda[A(y,\lambda/\beta) + R] - \ln \tilde{W}^{\beta}(y|x')\right)\right] \Big\} - R \\
= \; & \inf_{\{Q_{XX'Y}:\ F_Q(X,X')\leq 2R\}} \sup_{0\leq s\leq 1} \sup_{0\leq r\leq s} \inf_{\lambda>0} \Big[J_Q(X;X') + D(Q_X\|P) + \\
& \sum_{x,x'} Q_{XX'}(x,x') \sum_y Q_{Y|XX'}(y|xx') \left(\ln \frac{Q_{Y|XX'}(y|xx')}{W(y|x)} + \right. \\
& \left. r \ln \tilde{W}^{\beta}(y|x) + (s-r)\lambda[A(y,\lambda/\beta) + R] - s \ln \tilde{W}^{\beta}(y|x')\right) \Big] - R \\
\geq \; & \sup_{0\leq s\leq 1} \sup_{0\leq r\leq s} \inf_{\lambda>0} \inf_{\{Q_{XX'Y}:\ F_Q(X,X')\leq 2R\}} \Big[J_Q(X;X') + D(Q_X\|P) + \\
& \sum_{x,x'} Q_{XX'}(x,x') \sum_y Q_{Y|XX'}(y|xx') \times \\
& \left(\ln \frac{Q_{Y|XX'}(y|xx')\tilde{W}^{\beta r}(y|x)}{W(y|x)\tilde{W}^{\beta s}(y|x')e^{-\lambda(s-r)A(y,\lambda/\beta)}} - [1-\lambda(s-r)]R \right) \Big] \\
= \; & \sup_{0\leq s\leq 1} \sup_{0\leq r\leq s} \inf_{\lambda>0} \inf_{\{Q_{XX'}:\ F_Q(X,X')\leq 2R\}} \Big\{ J_Q(X;X') + D(Q_X\|P) +
\end{aligned}
$$

$$\sum_{x,x'} Q_{XX'}(x,x') \left(-\ln \left[\sum_y W(y|x) \cdot \frac{\tilde{W}^{\beta s}(y|x')}{\tilde{W}^{\beta r}(y|x) e^{\lambda(s-r)A(y,\lambda/\beta)}} \right] - [1-\lambda(s-r)]R \right) \Bigg\}$$

$$= \sup_{0\le s\le 1} \sup_{0\le r\le s} \inf_{\lambda>0} \inf_{\{Q_{XX'}:\ F_Q(X,X')\le 2R\}} \Bigg[J_Q(X;X') + D(Q_X\|P) + \sum_{x,x'} Q_{XX'}(x,x') \times$$

$$\left\{ -\ln \left(\sum_y W(y|x) \cdot \frac{\tilde{W}^{\beta s}(y|x')}{\tilde{W}^{\beta r}(y|x) \left[\sum_{\tilde{x}} P(\tilde{x}) \tilde{W}^{\beta/\lambda}(y|\tilde{x})\right]^{\lambda(s-r)}} \right) - [1-\lambda(s-r)]R \right\}.$$

Let us denote

$$G(x,x',\lambda,s,r) \stackrel{\Delta}{=} -\ln \left(\sum_y W(y|x) \cdot \frac{\tilde{W}^{\beta s}(y|x')}{\tilde{W}^{\beta r}(y|x) \left[\sum_{\tilde{x}} P(\tilde{x}) W^{1/\lambda}(y|\tilde{x})\right]^{\lambda(s-r)}} \right). \tag{33}$$

Then,

$$\begin{aligned}
E^g_{\rm trc}(R,P) \ \ge\ & \sup_{0\le s\le 1} \sup_{0\le r\le s} \inf_{\lambda\ge 0} \inf_{\{Q_{XX'}:\ F_Q(X,X')\le 2R\}} \Bigg[J_Q(X;X') + D(Q_X\|P) + \\
& \sum_{x,x'} Q_{XX'}(x,x') \left\{ G(x,x',\lambda,s,r) - [1-\lambda(s-r)]R \right\} \\
=\ & \sup_{0\le s\le 1} \sup_{0\le r\le s} \inf_{\lambda\ge 0} \inf_{\{Q_{XX'}:\ F_Q(X,X')\le 2R\}} \sum_{x,x'} Q_{XX'}(x,x') \left(\ln \frac{Q_{XX'}(x,x')}{P(x)P(x')} + \right. \\
& \left. G(x,x',\lambda,s,r) - [1-\lambda(s-r)]R \right) \\
=\ & \sup_{0\le s\le 1} \sup_{0\le r\le s} \inf_{\lambda\ge 0} \inf_{Q_{XX'}} \sup_{\zeta\ge 0} \sup_{\theta\ge 0} \sum_{x,x'} Q_{XX'}(x,x') \Bigg[\ln \frac{Q_{XX'}(x,x')}{P(x)P(x')} + \\
& \zeta \left(\ln \frac{Q_X(x)}{P(x)} - R \right) + \theta \left(\ln \frac{Q_{XX'}(x,x')}{P(x)P(x')} - 2R \right) + \\
& G(x,x',\lambda,s,r) - \{1-\lambda(s-r)\}R \Bigg] \\
=\ & \sup_{0\le s\le 1} \sup_{0\le r\le s} \inf_{\lambda\ge 0} \sup_{\zeta\ge 0} \sup_{\theta\ge 0} \inf_{Q_{XX'}} \sum_{x,x'} Q_{XX'}(x,x') \Bigg[\ln \frac{Q_{XX'}(x,x')}{P(x)P(x')} + \\
& \zeta \left(\ln \frac{Q_X(x)}{P(x)} - R \right) + \theta \left(\ln \frac{Q_{XX'}(x,x')}{P(x)P(x')} - 2R \right) + \\
& G(x,x',\lambda,s,r) - \{1-\lambda(s-r)\}R \Bigg] \\
=\ & \sup_{0\le s\le 1} \sup_{0\le r\le s} \inf_{\lambda\ge 0} \sup_{\zeta\ge 0} \sup_{\theta\ge 0} \inf_{Q_{XX'}} \sum_{x,x'} Q_{XX'}(x,x') \Bigg[(1+\theta) \ln \frac{Q_{XX'}(x,x')}{P(x)P(x')} + \\
& \zeta \ln \frac{Q_X(x)}{P(x)} + G(x,x',\lambda,s,r) - \{1-\lambda(s-r)+2\theta+\zeta\}R \Bigg] \\
=\ & \sup_{0\le s\le 1} \sup_{0\le r\le s} \inf_{\lambda\ge 0} \sup_{\zeta\ge 0} \sup_{\theta\ge 0} \inf_{Q_X} \sum_x Q_X(x) \Bigg[(1+\theta+\zeta) \ln \frac{Q_X(x)}{P(x)} + (1+\theta) \inf_{Q_{X'|X}} \\
& \left(\sum_{x'} Q_{X'|X}(x'|x) \ln \frac{Q_{X'|X}(x'|x)}{P(x')} + G(x,x',\lambda,s,r) - \{1-\lambda(s-r)+2\theta+\zeta\}R \right) \Bigg]
\end{aligned}$$

$$= \sup_{0\le s\le 1}\sup_{0\le r\le s}\inf_{\lambda\ge 0}\sup_{\zeta\ge 0}\sup_{\theta\ge 0}\inf_{Q_X}\sum_x Q_X(x)\left[(1+\theta+\zeta)\ln\frac{Q_X(x)}{P(x)}-\right.$$
$$\left.(1+\theta)\ln\left(\sum_{x'}P(x')e^{-G(x,x',\lambda,s,r)/(1+\theta)}\right)-\{1-\lambda(s-r)+2\theta+\zeta\}R\right]. \quad (34)$$

Denoting

$$T(x,\lambda,s,r,\theta)\stackrel{\Delta}{=}\ln\left(\sum_{x'}P(x')e^{-G(x,x',\lambda,s,r)/(1+\theta)}\right), \quad (35)$$

we finally, have

$$\begin{aligned}
E^g_{\rm trc}(R,P) &\ge \sup_{0\le s\le 1}\sup_{0\le r\le s}\inf_{\lambda\ge 0}\sup_{\zeta\ge 0}\sup_{\theta\ge 0}\left\{-(1+\theta+\zeta)\ln\left[\sum_x P(x)e^{(1+\theta)T(x,\lambda,s,r,\theta)/(1+\theta+\zeta)}\right]-\right.\\
&\quad\left.[1-\lambda(s-r)+2\theta+\zeta]R\right\}\\
&= \sup_{0\le s\le 1}\sup_{0\le r\le s}\inf_{\lambda\ge 0}\sup_{\zeta\ge 0}\sup_{\theta\ge 0}\left[-(1+\theta+\zeta)\ln\left\{\sum_x P(x)\left[\sum_{x'}P(x')\times\right.\right.\right.\\
&\quad\left.\left.\left.\left(\sum_y W(y|x)\cdot\frac{\tilde{W}^{\beta s}(y|x')}{\tilde{W}^{\beta r}(y|x)\left[\sum_{\tilde{x}}P(\tilde{x})tW^{\beta/\lambda}(y|\tilde{x})\right]^{\lambda(s-r)}}\right)^{1/(1+\theta)}\right]^{(1+\theta)/(1+\theta+\zeta)}\right\}-\right.\\
&\quad\left.[1-\lambda(s-r)+2\theta+\zeta]R\right]\\
&= \sup_{0\le\sigma\le\beta}\sup_{0\le\tau\le\beta-\sigma}\inf_{\lambda\ge 0}\sup_{\theta\ge 0}\sup_{\zeta\ge 1+\theta}\left[-\zeta\ln\left\{\sum_x P(x)\left[\sum_{x'}P(x')\times\right.\right.\right.\\
&\quad\left.\left.\left.\left(\sum_y W(y|x)\frac{\tilde{W}^{\sigma+\tau}(y|x')}{\tilde{W}^{\sigma}(y|x)\left[\sum_{\tilde{x}}P(\tilde{x})\tilde{W}^{1/\lambda}(y|\tilde{x})\right]^{\lambda\tau}}\right)^{1/(1+\theta)}\right]^{(1+\theta)/\zeta}\right\}-\right.\\
&\quad\left.(\zeta+\theta-\lambda\tau)R\right], \quad (36)
\end{aligned}$$

where in the last step, we have changed parameters according $\beta r\to\sigma$, $\beta s\to\sigma+\tau$, $\beta(s-r)=\tau$, and we have re–defined λ/β as λ, and $1+\theta+\zeta$ as ζ. This completes the proof of Theorem 1.

References

[1] R. Averbuch and N. Merhav, "Error exponents of typical random codes for source–channel coding," to appear in *Proc. ITW 2019*, Visby, Gotland. Sweden, 2019.

[2] R. Averbuch, N. Merhav, A. G. i Fábregas, and N. Weinberger, "Large deviations of random codes," in preparation.

[3] A. Barg and G. D. Forney, Jr., "Random codes: minimum distances and error exponents," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2568–2573, September 2002.

[4] G. Battail, "On random–like codes," *Proc. 4th Canadian Workshop on Information Theory*, pp. 76–94, Lac Delage, Quebec, Canada, May 1995.

[5] A. Bhatt, J.-T. Huang, Y.-H. Kim, J. J. Ryu, and P. Sen, "Variations on a theme by Liu, Cuff and Verdú: the power of posterior sampling," *Proc. 2018 Information Theory Workshop (ITW)*, Guangzhou, China, November 25–29, 2018.

[6] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, New York, 1968.

[7] J. Liu, P. Cuff and S. Verdú, "On α–decodability and α–liklihood decoder," *Proc. 55th Ann. Allerton Conf. Comm. Control Comput.*, Monticello, IL U.S.A., October 2017.

[8] N. Merhav, "The generalized stochastic likelihood decoder: random coding and expurgated bounds," *IEEE Trans. Inform. Theory*, vol. 63, no. 8, pp. 5039–5051, August 2017. See also correction at `https://arxiv.org/pdf/1707.03987.pdf`

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

[10] N. Merhav, "Error exponents of typical random trellis codes," submitted to *IEEE Trans. Inform. Theory*. Available on–line at `https://arxiv.org/pdf/1903.01120.pdf`

[11] N. Merhav, "Error exponents of typical random codes for the colored Gaussian channel," to appear in *IEEE Trans. Inform. Theory*, 2019. Available on–line at `https://arxiv.org/pdf/1812.06250.pdf`

[12] A. Nazari, *Error exponent for discrete memoryless multiple–access channels*, Ph.D. dissertation, Department of Electrical Engineering – Systems, the University of Michigan, 2011.

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

[14] J. Scarlett, A. Martinéz and A. G. i Fábregas, "The likelihood decoder: error exponents and mismatch," *Proc. 2015 IEEE International Symposium on Information Theory (ISIT 2015)*, pp. 86–90, Hong Kong, June 2015.

[15] E. C. Song, P. Cuff and H. V. Poor, "The likelihood encoder for lossy compression," *IEEE Trans. Inform. Theory*, vol. 62, no. 4, pp. 1836–1849, April 2016.

[16] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw-Hill, New York, 1979.

[17] M. H. Yassaee, M. R. Aref and A. Gohari, "A technique for deriving one–shot achievability results in network information theory," *Proc. 2013 IEEE International Symposium on Information Theory (ISIT 2013)*, pp. 1287–1291, July 2013. Also, available on–line at `http://arxiv.org/abs/1303.0696`.