Another Look at Expurgated Bounds and Their Statistical–Mechanical Interpretation*

Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
E–mail: merhav@ee.technion.ac.il

Abstract

We revisit the derivation of expurgated error exponents using a method of type class enumeration, which is inspired by statistical–mechanical methods, and which has already been used in the derivation of random coding exponents in several other scenarios. We compare our version of the expurgated bound to both the one by Gallager and the one by Csiszár, Körner and Marton (CKM). For expurgated ensembles of fixed composition codes over finite alphabets, our basic expurgated bound coincides with the CKM expurgated bound, which is in general tighter than Gallager's bound, but with equality for the optimum type class of codewords. Our method, however, extends beyond fixed composition codes and beyond finite alphabets, where it is natural to impose input constraints (e.g., power limitation). In such cases, the CKM expurgated bound may not apply directly, and our bound is in general tighter than Gallager's bound. In addition, while both the CKM and the Gallager expurgated bounds are based on Bhattacharyya bound for bounding the pairwise error probabilities, our bound allows the more general Chernoff distance measure, thus giving rise to additional improvement using the Chernoff parameter as a degree of freedom to be optimized.

Index Terms: Expurgated exponents, expurgated ensembles, Bhattacharyya distance, Chernoff distance, random energy model.

*This research was supported by the Israeli Science Foundation (ISF) grant no. 412/12.

1 Introduction

It is well known that the random coding exponent on the probability of error in channel coding can be improved, at low coding rates, by a process called *expurgation*, that results in the so called *expurgated exponent*, or the *expurgated bound*, which is a lower bound to the reliability function. The idea of expurgation, first introduced by Gallager [8, Section V], [9, Section 5.7] (see also [24, Section 3.3]), is that at low rates, the average error probability over the ensemble of codes, is dominated by bad randomly chosen codewords and not by the channel noise, therefore, by eliminating some of these codewords (while keeping the rate almost the same), an improved lower bound on the reliability function is obtained. The expurgated bound at zero rate is known to be tight, as it coincides, at this point, with the straight–line bound, which is an an upper bound on the reliability function [9, Section 5.8], [20], [21], [24, Sections 3.7, 3.8]. Omura [19] was the first to relate the expurgated exponent at low rates to distortion–rate functions, where the Bhattacharyya distance function plays the role of a distortion measure.

Several years later, Csiszár, Körner and Marton [3] derived, for finite alphabets, a different expurgated bound, henceforth referred to as the *CKM expurgated exponent*, as opposed to the *Gallager expurgated exponent* discussed above. While ref. [3] contains no details (it is an abstract only), the CKM expurgated exponent is mentioned in [1, eq. (7)] and some hints on its derivation can be found in [2, p. 185, Problem 17]. While the CKM expurgated exponent is equivalent to that of Gallager for the optimum channel input assignment [2, p. 193, Problem 23(b)], it turns out (as we will be shown below) that for a general input distribution, the CKM expurgated bound is larger (and hence tighter) than the Gallager expurgated bound. This is important whenever channel input constraints (e.g., power limitation) do not allow this optimum input distribution to be used. On the other hand, since the derivation [2, pp. 185–186, Problem 17 (hint)] of the CKM expurgated exponent relies strongly on the packing lemma [2, p. 162, Lemma 5.1], it is limited to finite input and output alphabets (as mentioned) and to fixed composition codes, as opposed to the Gallager expurgated exponent, whose derivation is carried out under more general conditions.

In this paper, our quest is to enjoy the best of both worlds: We use yet another analysis technique, which has already been used in several previous works in different scenarios [7], [10], [13], [14], [15, Chapters 6,7], [22], [23], where it has always yielded simplified and/or improved bounds

on error exponents. This technique, which is based on distance enumeration, or more generally, on type class enumeration, is inspired by the statistical–mechanical perspective on random coding, based on its analogy to the random energy model [18, Chapters 5, 6], which is a model of spin glasses with a high degree of disorder, invented by Derrida [4], [5], [6], and which is well known in the literature of statistical physics of magnetic materials. Our technique is applicable to channels with quite general input/output alphabets, it is not limited to fixed composition codes, and it allows the incorporation of channel input constraints, which are, of course, especially relevant when the channel input alphabet is continuous. In the special case of finite alphabets, our basic bound coincides with the CKM expurgated bound along the whole interesting range of rates, and hence is tighter, in general, than Gallager's expurgated exponent.

Furthermore, an additional improvement of our expurgated bound is obtained by observing that, instead of using the Bhattacharyya bound for the pairwise error probabilities (as is done in the derivations of both the Gallager- and the CKM expurgated exponents), it turns out that for our proposed form of the expurgated exponent, the pairwise error probabilities can more generally be bounded using the Chernoff distance measure, whose parameter is subjected to optimization.[1]

Finally, as mentioned above, our analysis technique is based on a statistical–mechanical point of view. This point of view naturally suggests a physical interpretation to the behavior of the expurgated exponent in the following sense: Similarly as in Gallager's and the CKM expurgated exponents, the graph of the new proposed expurgated exponent is curvy at low rates and becomes a straight line of slope -1 at the higher range of rates. It turns out that this passage from a curve to a straight line can be understood as a *phase transition* in the analogous statistical–mechanical system model – the random energy model. This point will be discussed as well.

The outline of the remaining part of this paper is as follows. In Section 2, we provide some background on the expurgated exponents of Gallager and Csiszár, Körner and Marton, as well as the relationship between them. In Section 3, we provide a few elementary observations that serve as a basis for our proposed derivation of the expurgated exponent. In Section 4, we present the derivation of the new proposed version of our expurgated error exponent for finite alphabets

[1] While Gallager's bound has a symmetry that guarantees that the optimum value of the Chernoff parameter is always 1/2 (in which case, the Chernoff distance coincides with the Bhattacharyya distance), this symmetry does not appear in the new proposed bound, and hence the optimum value of the Chernoff parameter is not necessarily 1/2.

and fixed composition codes. In Section 5, we outline the extension of this analysis to continuous alphabet channels. Finally, in Section 6, we discuss the statistical–mechanical perspective of our analysis.

2 Background

Consider a discrete memoryless channel (DMC), defined by the single–letter transition probability functions[2] $P = \{p(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input alphabet and the output alphabet, respectively. Let $Q = \{q(x),\ x \in \mathcal{X}\}$ be a probability function on the input alphabet $\mathcal{X}$.

Gallager's random coding error exponent function is a well known lower bound on the reliability function of the DMC [8], [9, Section 5.6], [24, Section 3.2]. It is given by

$$E_r(R) = \sup_{0 \le \rho \le 1} \sup_{Q} [E_0(\rho, Q) - \rho R] \tag{1}$$

where

$$E_0(\rho, Q) = -\ln \left(\sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} q(x) p(y|x)^{1/(1+\rho)} \right]^{1+\rho} \right), \tag{2}$$

and where here and throughout the sequel, it is understood that for continuous alphabets, summations are replaced by integrals. This bound is obtained by analyzing the exponential rate of the average error probability associated with a randomly chosen code $\mathcal{C}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$, $M = e^{nR}$, R being the coding rate and $\boldsymbol{x}_m \in \mathcal{X}^n$ being the codeword associated message number $m \in \{1, \ldots, M\}$, where each component of each codeword is selected independently at random under Q.

At low rates, this lower bound on the reliability function can be improved by expurgating the randomly chosen code. This expurgation is accomplished by discarding the 'bad' half of the codebook, namely, the half of codewords whose conditional error probabilities

$$P_{e|m} = \Pr\{\text{error}|\text{message } m \text{ sent}\}$$

are the largest under maximum likelihood (ML) decoding. Gallager's expurgated exponent function [8], [9, Section 5.7] [24, Section 3.3] is given by

$$E_{ex}(R) = \sup_{\rho \ge 1} \sup_{Q} [E_x(\rho, Q) - \rho R] \tag{3}$$

[2]Here and throughout the sequel, "probability function" is a common name for a probability mass function in the discrete alphabet case and a probability density function in the continuous alphabet case.

where

$$E_x(\rho, Q) = -\rho \ln \left(\sum_{x,x' \in \mathcal{X}} q(x)q(x') \left[\sum_{y \in \mathcal{Y}} \sqrt{p(y|x)p(y|x')} \right]^{1/\rho} \right). \tag{4}$$

Improvement over $E_r(R)$ is accomplished whenever the coding rate R is small enough such that the supremum in eq. (3) is achieved (or approached) by values of ρ that are strictly larger than 1, as otherwise for $\rho = 1$, we have $E_x(1, Q) \equiv E_0(1, Q)$.

In [3] (see also [2, p. 185, Problem 17] for details), the following version of the expurgated exponent was presented by Csiszár, Körner and Marton (CKM) for channels with finite input and output alphabets:

$$E_{ex}(R) = \sup_{Q} \inf_{\hat{Q}_{XX'} \in \mathcal{A}(R,Q)} [I(X;X') + \boldsymbol{E} d_B(X, X')] - R, \tag{5}$$

where $\hat{Q}_{XX'}$ is a generic joint probability mass function over $\mathcal{X}^2$, that governs both the mutual information and the expectation in the square brackets of eq. (5),

$$\mathcal{A}(R, Q) = \{\hat{Q}_{XX'} : \ \hat{Q}_X = \hat{Q}_{X'} = Q, \ I(X;X') \leq R\},$$

and $d_B(\cdot, \cdot)$ is the *Bhattacharyya distance function*, defined by

$$d_B(x, x') = -\ln \left[\sum_{y \in \mathcal{Y}} \sqrt{p(y|x)p(y|x')} \right]. \tag{6}$$

In [2, p. 193, Problem 23b] it is asserted that the right–hand sides of eqs. (3) and (5) are equivalent, thus justifying the common notation $E_{ex}(R)$ for both expressions. Hereafter, to avoid confusion between the Gallager and the CKM expurgated exponents, we will deviate from the customary notation used above, and re–define the notation $E_G(\rho, Q)$ for $E_x(\rho, Q)$ (where the subscript G stands for "Gallager"), and accordingly

$$\mathcal{E}_G(R, Q) = \sup_{\rho \geq 1} [E_G(\rho, Q) - \rho R], \tag{7}$$

thus, $E_{ex}(R) = \sup_Q \mathcal{E}_G(R, Q)$. Similarly, we will denote

$$\mathcal{E}_{CKM}(R, Q) = \inf_{\hat{Q}_{XX'} \in \mathcal{A}(R,Q)} [I(X;X') + \boldsymbol{E} d_B(X, X')] - R, \tag{8}$$

thus, $E_{ex}(R) = \sup_Q \mathcal{E}_{CKM}(R, Q)$.

While $\sup_Q \mathcal{E}_G(R,Q) = \sup_Q \mathcal{E}_{CKM}(R,Q)$ as mentioned above, it turns out that for a general choice of Q, the functions $\mathcal{E}_G(R,Q)$ and $\mathcal{E}_{CKM}(R,Q)$ may differ. In fact, as we shall see shortly

$$\mathcal{E}_{CKM}(R,Q) \geq \mathcal{E}_G(R,Q) \tag{9}$$

for an arbitrary input assignment Q. This is an important point since the optimum input assignment Q^*, that achieves $E_{ex}(R)$, might be forbidden in the presence of channel input constraints (e.g., power limitation), and so, in such a case, the CKM expurgated exponent may be better than the Gallager expurgated exponent. On the other hand, there are two advantages to the Gallager expurgated exponent relative to the CKM expurgated exponent. The first is that, unlike the case of the CKM bound, its derivation is not sensitive to the assumption of finite alphabets and fixed composition codes.[3] The second advantage is that the numerical calculation of $\mathcal{E}_G(R,Q)$ requires optimization over one parameter only (the parameter ρ), whereas the calculation of $\mathcal{E}_{CKM}(R,Q)$ seems (at least in its present form) to require optimization over the entire joint distribution $\hat{Q}_{XX'}$ (which means many parameters for a large input alphabet) and moreover, this optimization is subjected to complicated constraints (defined by $\mathcal{A}(R,Q)$).

3 Some Preliminary Observations

Before presenting the proposed alternative derivation of our expurgated exponent, we pause to offer a few preliminary observations that would hopefully help to compare $\mathcal{E}_G(R,Q)$ and $\mathcal{E}_{CKM}(R,Q)$ and to understand the relationships between them, as well as their relation to that of the new bound to be derived. In particular, our first task is to transform the expression of $\mathcal{E}_{CKM}(R,Q)$ to a form that has the same ingredients as those of $\mathcal{E}_G(R,Q)$.

We first define the function

$$D_Q(R) = \min_{\hat{Q}_{XX'} \in \mathcal{A}(R,Q)} \boldsymbol{E}\{d_B(X;X')\}. \tag{10}$$

Intuitively, the function $D_Q(R)$ is the distortion–rate function of a "source" Q (designated by the random variable X) with respect to (w.r.t.) the Bhattacharyya distortion measure $d_B(\cdot,\cdot)$, subject to the additional constraint that the "reproduction variable" X' has the same probability

[3] In fact, in the case of a continuous input alphabet, the notion of fixed composition codes does not really exist altogether.

distribution Q as the "source." It is easy to see now that

$$\inf_{\hat{Q}_{XX'}\in\mathcal{A}(R,Q)}[I(X;X')+\boldsymbol{E}d_B(X,X')]=\begin{cases} D_Q(R)+R & R\le R_1\\ D_Q(R_1)+R_1 & R>R_1\end{cases} \tag{11}$$

where R_1 is $I(X;X')$ for the optimum $\hat{Q}_{XX'}$ that minimizes $[I(X;X')+\boldsymbol{E}d_B(X,X')]$ across $\mathcal{A}(\infty,Q)$, or equivalently, R_1 is the rate R at which $D'_Q(R)=-1$, $D'_Q(R)$ being the derivative of $D_Q(R)$ w.r.t. R. Thus, we obtain

$$\mathcal{E}_{CKM}(R,Q)=\begin{cases} D_Q(R) & R\le R_1\\ D_Q(R_1)+R_1-R & R_1<R<D_Q(R_1)+R_1\\ 0 & R>D_Q(R_1)+R_1\end{cases} \tag{12}$$

where we note that the first line is intimately related to [2, p. 194, Problem 24]. We observe then that at low rates, $\mathcal{E}_{CKM}(R,Q)$ has a curvy part given by $D_Q(R)$, and for high rates it is given by the straight line of slope -1 that is tangential to the curve $D_Q(R)$.

Let us now take a closer look at the distortion–rate function $D_Q(R)$, which is the inverse of the rate–distortion function $R_Q(D)$, defined similarly, and again with the additional constraint $Q_{X'}=Q$. This rate–distortion function has the following parametric representation [17, eq. (13)]:

$$R_Q(D)=-\inf_{s\ge 0}\left[sD+\sum_{x\in\mathcal{X}}q(x)\ln\left(\sum_{x'\in\mathcal{X}}q(x')e^{-sd_B(x,x')}\right)\right], \tag{13}$$

where the minimizing s is interpreted as the negative local slope of the function $R_Q(D)$, i.e., $s^*=-R'_Q(D)$, s^* being the minimizer of the r.h.s. This function can easily be inverted, similarly as in [16, eqs. (15)–(20)], to obtain

$$\begin{aligned} D_Q(R) &= -\inf_{s\ge 0}\frac{1}{s}\left[R+\sum_{x\in\mathcal{X}}q(x)\ln\left(\sum_{x'\in\mathcal{X}}q(x')e^{-sd_B(x,x')}\right)\right] &(14)\\ &= \sup_{\rho\ge 0}\left[-\rho\sum_{x\in\mathcal{X}}q(x)\ln\left(\sum_{x'\in\mathcal{X}}q(x')e^{-d_B(x,x')/\rho}\right)-\rho R\right], &(15)\end{aligned}$$

where the second line follows from the first simply by changing the variable s to the variable $\rho=1/s$. Thus, the maximizing ρ is the negative local slope of the function $D_Q(R)$. It follows that in the curvy part of $\mathcal{E}_{CKM}(R,Q)$, where the slope of $D_Q(R)$ is smaller than -1, the maximizing ρ is larger than 1. Thus, the maximization in the last expression of $D_Q(R)$ can be confined to the range $[1,\infty)$, i.e., for $R\le R_1$

$$\mathcal{E}_{CKM}(R,Q) = \sup_{\rho\ge 1}\left[-\rho\sum_{x\in\mathcal{X}}q(x)\ln\left(\sum_{x'\in\mathcal{X}}q(x')e^{-d_B(x,x')/\rho}\right)-\rho R\right]$$

$$= \sup_{\rho \geq 1} \left\{ -\rho \sum_{x \in \mathcal{X}} q(x) \ln \left(\sum_{x' \in \mathcal{X}} q(x') \left[\sum_{y \in \mathcal{Y}} \sqrt{p(y|x)p(y|x')} \right]^{1/\rho} \right) - \rho R \right\}. \quad (16)$$

and of course, for $R \in [R_1, R_1 + D_Q(R_1)]$ we use the same expression, setting $\rho = 1$. This should now be compared with Gallager's expression

$$\mathcal{E}_G(R, Q) = \sup_{\rho \geq 1} \left\{ -\rho \ln \left(\sum_{x, x' \in \mathcal{X}} q(x) q(x') \left[\sum_{y \in \mathcal{Y}} \sqrt{p(y|x)p(y|x')} \right]^{1/\rho} \right) - \rho R \right\}. \quad (17)$$

As can be seen, the only difference between the two expressions is that in $\mathcal{E}_{CKM}(R, Q)$, the averaging over x is external to the logarithmic function, whereas in $\mathcal{E}_G(R, Q)$ it is internal to the logarithmic function. Thus, Jensen's inequality guarantees that $\mathcal{E}_{CKM}(R, Q) \geq \mathcal{E}_G(R, Q)$, and since the logarithmic function is strictly concave, the inequality is strict for every finite ρ (which means $R > 0$), unless $\sum_{x' \in \mathcal{X}} q(x') e^{-d_B(x,x')/\rho}$ happens to be independent of x, which is the case when either Q and P exhibit enough symmetry, or when Q is chosen to be the optimum distribution [2, p. 193, Problem 23b, hint (iii)].

Our second preliminary observation is the following. The derivation of Gallager's expurgated exponent begins from the union bound on the pairwise error probabilities, which in turn are all upper bounded by the Bhattacharyya bound, i.e., eq. (5.7.3) in [9] reads

$$P_{e|m} \leq \sum_{m' \neq m} \sum_{\boldsymbol{y}} \sqrt{p(\boldsymbol{y}|\boldsymbol{x}_m) p(\boldsymbol{y}|\boldsymbol{x}_{m'})}, \quad (18)$$

where $\boldsymbol{y} \in \mathcal{Y}^n$ designates the channel output vector. One might suspect that a better result can probably be obtained by considering, more generally, the Chernoff bound

$$P_{e|m} \leq \sum_{m' \neq m} \sum_{\boldsymbol{y}} p^s(\boldsymbol{y}|\boldsymbol{x}_{m'}) p^{1-s}(\boldsymbol{y}|\boldsymbol{x}_m), \qquad 0 \leq s \leq 1, \quad (19)$$

where the Chernoff parameter s is subjected to optimization (in addition to the parameter ρ). After carrying out the derivation similarly as in [9, Section 5.7], one would obtain a similar expression as in $\mathcal{E}_G(R, Q)$, except that the Bhattacharyya distance function is replaced, more generally, by the Chernoff distance function

$$d_s(x, x') = -\ln \left[\sum_{y \in \mathcal{Y}} p^{1-s}(y|x) p^s(y|x') \right]. \quad (20)$$

Thus, $E_G(\rho, Q)$ would be replaced by

$$E_G(\rho, s, Q) = -\rho \ln \left(\sum_{x,x' \in \mathcal{X}} q(x)q(x') \left[\sum_{y \in \mathcal{Y}} p^{1-s}(y|x)p^s(y|x') \right]^{1/\rho} \right), \tag{21}$$

and the best choice of s would be the one that maximizes $E_G(\rho, s, Q)$. However, it is easy to see that $E_G(\rho, s, Q)$ is concave in s and that $E_G(\rho, s, Q) = E_G(\rho, 1-s, Q)$ since x and x' play symmetric roles in the expression of $E_G(\rho, s, Q)$. Thus, the maximizing s is obviously $s^* = 1/2$, which brings us back to the Bhattacharyya distance, and confirming that there is nothing to gain from the optimization over s beyond Gallager's expurgated bound.

This is not the case, however, when it comes to the CKM expurgated bound. In particular, Csiszár and Körner also begin from the union–Bhattacharyya bound (see [2, p. 186, top]), and an extension of their derivation would yield the same expression as (16), but again, with the Bhattacharyya distance $d_B(x, x')$ (or $d_{1/2}(x, x')$) being replaced by the more general Chernoff distance $d_s(x, x')$. However, here x and x' do *not* have symmetric roles and hence the bound is not necessarily optimized at $s = 1/2$. Indeed, it is easy to study a simple example of a binary non–symmetric channel and see that the derivative of the function

$$E(\rho, s, Q) = -\rho \sum_{x \in \mathcal{X}} q(x) \ln \left(\sum_{x' \in \mathcal{X}} q(x') \left[\sum_{y \in \mathcal{Y}} p^{1-s}(y|x)p^s(y|x') \right]^{1/\rho} \right) \tag{22}$$

with respect to s does not vanish at $s = 1/2$ unless Q is symmetric (see also Example 1 below, at the end of this section).

To summarize, we observe that the CKM expurgated bound is not only better, in general, than the Gallager expurgated bound, but moreover, it provides even further room for improvement in the optimization over s, in addition to the optimization over ρ. Confining the framework to finite alphabets and fixed composition codes, this gives rise to the following coding theorem.

Theorem 1 *For an arbitrary DMC, there exist a sequence of codes $\{\mathcal{C}_n\}_{n \geq 1}$ of rate R and composition Q,*[4] *for which the error exponent associated with the maximum error probability is at least as large as*

$$\mathcal{E}(R, Q) = \sup_{\rho \geq 1} \sup_{0 \leq s \leq 1} [E(\rho, s, Q) - \rho R] \tag{23}$$

[4] A sequence of codes with composition Q means a sequence of fixed composition codes, where the common empirical distribution of all codewords tends to Q as $n \to \infty$.

where $E(\rho, s, Q)$ is defined as in eq. (22).

Example 1 – binary input, binary output channels. We have compared numerically the three expurgated exponents for various combinations of P and Q associated with binary input, binary output channels. As a representative example, we have computed $E_G(1, Q)$, $E(1, 1/2, Q)$ and $\max_{0 \le s \le 1} E(1, s, Q)$, for the binary channel P defined by $p(0|0) = p(1|0) = 0.5$, $p(0|1) = 1 - p(1|1) = 10^{-10}$, along with the input assignment Q given by $q(1) = 1 - q(0) = 0.1$. The results are $E_G(1, Q) = 0.0542$, $E(1, 1/2, Q) = 0.0574$, and $\max_{0 \le s \le 1} E(1, s, Q) = 0.0596$, which is achieved at $s^* \approx 0.76$. This means that in the range of high rates, we have

$$\mathcal{E}_G(R, Q) = 0.0542 - R \tag{24}$$

$$\mathcal{E}_{CKM}(R, Q) = 0.0574 - R \tag{25}$$

$$\mathcal{E}(R, Q) = 0.0596 - R. \tag{26}$$

Thus, numerical evidence indeed supports the fact that there are gaps between the three expurgated exponents, at least for some combinations of channels and input assignments.

4 New Derivation of the Expurgated Exponent

Equipped with the background of Section 2 and the observations offered in Section 3, we next proceed to the derivation of the new version of the expurgated bound (i.e., prove Theorem 1), but in a manner that does not rely on the packing lemma and hence is not sensitive to the assumptions of fixed composition codes and finite alphabets. We will assume finite alphabets only for the simplicity of the exposition and for the sake convenience, but it should be understood that our analysis has a natural extension to continuous alphabets (along with channel input constraints), and we will outline this extension in Section 5.

Following the discussion in Section 3, we begin with the following upper bound on the conditional probability of error

$$P_{e|m} \le \sum_{m' \ne m} \sum_{\boldsymbol{y}} p^s(\boldsymbol{y}|\boldsymbol{x}_{m'}) p^{1-s}(\boldsymbol{y}|\boldsymbol{x}_m), \qquad 0 \le s \le 1. \tag{27}$$

Now, following the same rationale as in [9, Section 5.7] and [24, Section 3.3], we argue the following: There exists a codebook $\mathcal{C}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ of $M = e^{nR}$ codewords such that for every $\rho > 0$ and

all $1 \leq m \leq M$

$$P_{e|m} \leq \left[2\overline{P_{e|m}^{1/\rho}}\right]^{\rho} \leq 2^{\rho} \left[\boldsymbol{E}\left(\sum_{m' \neq m} \sum_{\boldsymbol{y}} p^{s}(\boldsymbol{y}|\boldsymbol{X}_{m'}) p^{1-s}(\boldsymbol{y}|\boldsymbol{X}_{m})\right)^{1/\rho}\right]^{\rho} \stackrel{\Delta}{=} 2^{\rho} A_n(R,\rho), \tag{28}$$

where the expectation operator is taken w.r.t. the randomness of the codewords $\{\boldsymbol{X}_m\}$, which are selected independently at random according to the uniform distribution over the type class T_Q, that is, the set of all sequences whose empirical distribution is (as close as possible to) Q.

For the purpose of further bounding $A_n(R,\rho)$, the next step in both [9] and [24] is to use the inequality $[\sum_{m'} a_{m'}]^{1/\rho} \leq \sum_{m'} a_{m'}^{1/\rho}$, which holds for every $\rho \geq 1$, and then to apply the expectation operator on each term of the corresponding sum separately. This is a step which simplifies the derivation to a large extent, but at the possible price of losing exponential tightness of the resulting bound. Instead, in our derivation, we will use another approach, which yields an exponentially tight bound. Defining

$$d_s(x,x') = -\ln\left[\sum_{y} p^{1-s}(y|x) p^{s}(y|x')\right] \tag{29}$$

we have, due to the memorylessness of the channel,

$$\sum_{\boldsymbol{y}} p^{1-s}(\boldsymbol{y}|\boldsymbol{x}_m) p^{s}(\boldsymbol{y}|\boldsymbol{x}_{m'}) = e^{-\sum_{i=1}^{n} d_s(x_{m,i},x_{m',i})} \stackrel{\Delta}{=} e^{-d_s(\boldsymbol{x}_m,\boldsymbol{x}_{m'})}, \tag{30}$$

where $x_{m,i}$ is the i–th component of the codeword $\boldsymbol{x}_m$. Let $N_m(\hat{Q}_{XX'})$ be the number of codewords $\{\boldsymbol{x}_{m'}\}$ that, together with $\boldsymbol{x}_m$, fall in the joint type class corresponding to the joint empirical distribution $\hat{Q}_{XX'}$, whose both marginals must agree with Q (as they are both empirical distributions of codewords). Then, we have

$$\begin{aligned}
A_n(R,\rho) &= \left[\boldsymbol{E}\left(\sum_{\hat{Q}_{XX'}} N_m(\hat{Q}_{XX'}) \exp\{-n\boldsymbol{E}d_s(X,X')\}\right)^{1/\rho}\right]^{\rho} \\
&\doteq \left[\boldsymbol{E}\left(\max_{\hat{Q}_{XX'}} N_m(\hat{Q}_{XX'}) \exp\{-n\boldsymbol{E}d_s(X,X')\}\right)^{1/\rho}\right]^{\rho} \\
&= \left[\boldsymbol{E} \max_{\hat{Q}_{XX'}} [N_m(\hat{Q}_{XX'})]^{1/\rho} \exp\{-n\boldsymbol{E}d_s(X,X')/\rho\}\right]^{\rho} \\
&\doteq \left[\boldsymbol{E} \sum_{\hat{Q}_{XX'}} [N_m(\hat{Q}_{XX'})]^{1/\rho} \exp\{-n\boldsymbol{E}d_s(X,X')/\rho\}\right]^{\rho}
\end{aligned}$$

$$
\begin{aligned}
&= \left[\sum_{\hat{Q}_{XX'}} \boldsymbol{E}\left\{[N_m(\hat{Q}_{XX'})]^{1/\rho}\right\} \cdot \exp\{-n\boldsymbol{E}d_s(X,X')/\rho\}\right]^{\rho} \\
&\doteq \left[\max_{\hat{Q}_{XX'}} \boldsymbol{E}\left\{[N_m(\hat{Q}_{XX'})]^{1/\rho}\right\} \cdot \exp\{-n\boldsymbol{E}d_s(X,X')/\rho\}\right]^{\rho} \\
&\doteq \max_{\hat{Q}_{XX'}} \left(\boldsymbol{E}\left\{[N_m(\hat{Q}_{XX'})]^{1/\rho}\right\}\right)^{\rho} \cdot \exp\{-n\boldsymbol{E}d_s(X,X')\}, \qquad (31)
\end{aligned}
$$

where the notation $\doteq$ designates equivalence in the exponential scale (i.e., $a_n \doteq b_n$ means that $\frac{1}{n}\ln\frac{a_n}{b_n} \to 0$ as $n \to \infty$), and where the expectation at the exponent is w.r.t. $\hat{Q}_{XX'}$. Now, similarly as in [12, p. 4444, eq. (34)], we have

$$
\boldsymbol{E}\left\{[N_m(\hat{Q}_{XX'})]^{1/\rho}\right\} \doteq \begin{cases} \exp\{n[R - I(X;X')]\} & R < I(X;X') \\ \exp\{n[R - I(X;X')]/\rho\} & R \geq I(X;X') \end{cases} \qquad (32)
$$

where $I(X;X')$ is the mutual information between X and X' associated with $\hat{Q}_{XX'}$. This result follows from the fact that given $\boldsymbol{X}_m = \boldsymbol{x}_m$, $N_m(\hat{Q}_{XX'})$ is the sum of $e^{nR} - 1$ binary independent random–variables,

$$
U_{m'} = 1\{(\boldsymbol{x}_m, \boldsymbol{X}_{m'}) \text{ have empirical joint distribution } \hat{Q}_{XX'}\}, \quad m' \neq m, \qquad (33)
$$

whose expectations are all of the exponential order of $e^{-nI(X;X')}$. Upon taking into account all the possible empirical distributions $\{\hat{Q}_{XX'}\}$, we readily obtain

$$
A_n(R,\rho) \doteq e^{-n\min\{E_1(R), E_2(R,\rho)\}}, \qquad (34)
$$

where

$$
E_1(R,\rho) = \min_{\hat{Q}_{X'|X}:\ I(X;X') \geq R} [Ed_s(X,X') + \rho I(X;X')] - \rho R \qquad (35)
$$

and

$$
E_2(R) = \min_{Q_{X'|X}:\ I(X;X') \leq R} [Ed_s(X,X') + I(X;X')] - R = \sup_{\rho \geq 1}[E(\rho, s, Q) - \rho R], \qquad (36)
$$

where the second equality is obtained similarly as in the derivation of eq. (16), but with the Bhattacharyya distortion measure being replaced by $d_s(\cdot,\cdot)$. It remains to show that $E_1(R,\rho)$, for the optimum choice of ρ, is never smaller than $\mathcal{E}_{CKM}(R,Q)$. For a given s, let $R_Q(D)$ be the rate–distortion function of X w.r.t. the distortion measure $\{d_s(x,x')\}$ subject to the constraint that $Q_{X'} = Q$. Let D_ρ be the distortion level at which $R'_Q(D) = -1/\rho$, where $R'_Q(\cdot)$ is the derivative of

$R_Q(\cdot)$. Also, $D_Q(R)$ will denote the corresponding distortion–rate function, which is the inverse of $R_Q(D)$. Then $E_1(R,\rho)$ admits the following expressions:

$$E_1(R,\rho) = \begin{cases} D_\rho + \rho[R_Q(D_\rho) - R] & R \le R_Q(D_\rho) \\ D_Q(R) & R \ge R_Q(D_\rho) \end{cases} \tag{37}$$

As the straight line $D_\rho + \rho[R_Q(D_\rho) - R]$ is tangential to (and below) the convex function $D_Q(R)$, the best choice of ρ is to take the limit $\rho \to \infty$. But $E_1(R,\infty) = D_Q(R)$ for all R (as $R_Q(D_\infty) = 0$), which is in turn at least as large as $E_2(R) = \sup_{\rho \ge 1}[E(\rho, s, Q) - \rho R]$ for all R, and strictly so in the linear part of the latter function.

Thus, for a given s, there exists a sequence of codes for which the exponent of the maximum probability of error is dominated by $\sup_{\rho \ge 1}[E(\rho, s, Q) - \rho R]$. Upon maximization over s, this yields $\mathcal{E}(R, Q)$, as asserted in Theorem 1.

5 Beyond Finite Alphabets and Fixed Composition Codes

In Section 4, we have assumed finite alphabets and fixed composition codes, mainly for the simplicity of the exposition and for the purpose of comparison with the CKM expurgated exponent. However, as we have mentioned already, the analysis in Section 4 is not really sensitive to these assumptions.

The heart of the analysis in Section 4 is around equations (31) and (32), and therefore, the main issue in the desired extension is to adapt this part of the analysis to continuous alphabets. Consider now the case where $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and then $q(x)$ and $p(y|x)$ are probability density functions. Let δ be an arbitrarily small positive real. Then,

$$\sum_{m' \ne m} e^{-d_s(\boldsymbol{x}_m, \boldsymbol{x}_{m'})} \le \sum_{k=0}^{\infty} e^{-nk\delta} N_m(k), \tag{38}$$

where

$$N_m(k) = \sum_{m' \ne m} 1\left\{nk\delta \le d_s(\boldsymbol{x}_m, \boldsymbol{x}_{m'}) < n(k+1)\delta)\right\}, \qquad k = 0, 1, 2, \ldots \tag{39}$$

Let us assume now that the ensemble of codes is defined such that $d(\boldsymbol{x}_m, \boldsymbol{x}_{m'})$ cannot exceed $nD_{\max}$, where $D_{max} < \infty$ is a constant that does not depend on n, which is normally the case when the codewords must comply with input constraints. Then using a similar technique as in eq. (31), we now obtain

$$A_n(R,\rho) \stackrel{\cdot}{\le} \sup_{k \ge 0} \left(\boldsymbol{E}\{[N_m(k)]^{1/\rho}\}\right)^\rho \cdot e^{-nk\delta}, \tag{40}$$

where the notation $\dot{\leq}$ denotes inequality in the exponential scale (more formally, $a_n \dot{\leq} b_n$ means $\limsup_{n\to\infty} \frac{1}{n} \ln \frac{a_n}{b_n} \leq 0$). The key issue is now to assess the exponential rate of the expectation of the binary random variable,

$$U_{m'} = 1\left\{nk\delta \leq d_s(\boldsymbol{x}_m, \boldsymbol{X}_{m'}) < n(k+1)\delta\right\}, \tag{41}$$

for a given $\boldsymbol{x}_m$, namely, to find the exponent of $\Pr\{nk\delta \leq d_s(\boldsymbol{x}_m, \boldsymbol{X}_{m'}) < n(k+1)\delta\}$. This can be done using standard large deviations techniques, like the Chernoff bound. Let $R(k\delta)$ denote the large deviations rate function of this probability (which depends, of course, on $\boldsymbol{x}_m$, but it would be convenient to define the ensemble such that this rate function will be the same for all m). Then, as in eq. (32), we then have

$$\boldsymbol{E}\{[N_m(k)]^{1/\rho}\} \dot{=} \begin{cases} \exp\{n[R - R(k\delta)]\} & R \leq R(k\delta) \\ \exp\{n[R - R(k\delta)]/\rho\} & R > R(k\delta) \end{cases} \tag{42}$$

Now, similarly as in Section 4, $A_n(R,\rho)$ is dominated by $\min\{E_1(R,\rho,\delta), E_2(R,\delta)\}$, where

$$E_1(R,\rho,\delta) = \inf_{k:\ R(k\delta)\geq R} [k\delta + \rho R(k\delta)] - \rho R, \tag{43}$$

and

$$E_2(R,\delta) = \inf_{k:\ R(k\delta)\leq R} [k\delta + R(k\delta)] - R, \tag{44}$$

Upon taking the limit $\delta \to 0$, these become

$$E_1(R,\rho) = \inf_{D:\ R(D)\geq R} [D + \rho R(D)] - \rho R, \tag{45}$$

and

$$E_2(R) = \inf_{D:\ R(D)\leq R} [D + R(D)] - R. \tag{46}$$

The remaining details depend, of course, on the form of the large deviations rate function $R(D)$, which in turn depends strongly on the input assignment and the channel.

Example 2 – the Gaussian channel. Consider the memoryless additive Gaussian channel $Y = X + Z$, where Z is a zero–mean Gaussian random variable with variance σ^2, independent of X. Let $q(\boldsymbol{x})$ be the uniform distribution over the surface of the n–dimensional sphere with radius $\sqrt{nS}$. In this case, the Chernoff distance is maximized at $s^* = 1/2$, where it agrees with the Bhattacharyya distance $d_B(x,x') = (x-x')^2/8\sigma^2$. It is not difficult to show (e.g., using the methods of [11]) that

$$R(D) = \frac{1}{2} \ln \left[\frac{S}{8\sigma^2 D(1 - 2\sigma^2 D/S)} \right], \tag{47}$$

which has the interpretation of the rate–distortion function of the Gaussian source with variance S w.r.t. Bhattacharyya distortion measure with the additional constraint that reproduction variable X' is also Gaussian, zero–mean and with variance S. The corresponding distortion–rate function (which is the inverse of $R(D)$) is given by

$$D(R) = \frac{S(1 - \sqrt{1 - e^{-2R}})}{4\sigma^2}, \tag{48}$$

which is also the curvy part of the corresponding expurgated exponent. The linear part is again the tangential straight line with slope -1.

6 The Statistical–Mechanical Perspective

Let us take another look at the central expression that was handled in Sections 4 and 5, namely, on the summation

$$Z = \sum_{m' \neq m} e^{-d_s(\boldsymbol{x}_m, \boldsymbol{x}_{m'})}, \tag{49}$$

From the viewpoint of statistical physics, this can be interpreted as the partition function of a physical system, where for a fixed $\boldsymbol{x}_m$, the various configurations (microstates) are $\{\boldsymbol{x}_{m'}\}_{m' \neq m}$ and the Hamiltonian (energy function) is given by (or proportional[5] to) $d_s(\boldsymbol{x}_m, \boldsymbol{x}_{m'})$. If the correct codeword $\boldsymbol{x}_m$ is given and the remaining codewords are considered independent and random, thus denoted $\{\boldsymbol{X}_{m'}\}$, then the various "configurational energies" $\{d_s(\boldsymbol{x}_m, \boldsymbol{X}_{m'})\}$ are also independent random variables. As explained in [18, Chapters 5, 6] (see also [15, Chapters 6, 7] and references therein), this setting is analogous to the random energy model (REM) in the literature of statistical physics of magnetic materials. The REM was invented by Derrida [4], [5], [6] as a model of extremely disordered spin glasses. This model is not realistic, but it is exactly solvable and it exhibits a phase transition: Below a certain critical temperature, the partition function becomes dominated by a sub–exponential number of configurations, which means that the system freezes in the sense that its entropy vanishes in the thermodynamic limit. This combination of freezing and quenched disorder resembles the behavior of a glass, and so, this low temperature phase of zero entropy is called the *glassy phase*.[6] Above the critical temperature, the partition function is dominated by an

[5]To enhance the analogy with physics, it is instructive to consider a parametric family of channels, $p_\beta(y|x) \propto [p(y|x)]^\beta$, where β is a parameter that controls the 'quality' of the channel (e.g., the SNR in the case of the Gaussian channel), whose physical meaning is inverse temperature. In this case, $d_s(\boldsymbol{x}_m, \boldsymbol{x}_{m'})$ of the channel pertaining to $\beta = 1$ would be multiplied by β, similarly as in ordinary partition functions.

[6]In physics, it typically occurs as a result of a process of rapid cooling.

exponential number of configurations, and so, its entropy is positive. This high temperature phase is called the *paramagnetic phase.*

In the derivations of Sections 4 and 5, the curvy part of the graph of $\mathcal{E}(R,Q)$ corresponds to the glassy phase of the REM associated with (49), because the dominant contribution to $A_n(R,\rho)$ is due to a subexponential number ($N_m(\hat{Q}_{XX'})$ or $N_m(k)$) of codewords whose distance from $\boldsymbol{x}_m$ is about $nD_Q(R)$. The straight–line part of $\mathcal{E}(R,Q)$, on the other hand, corresponds to the paramagnetic phase, where about $e^{n[R-R_1]}$ incorrect codewords at distance $nD_Q(R_1)$ dictate the behavior. Thus, the passage between the curvy part and the straight–line part, at $R=R_1$ is interpreted as a glassy phase transition.

In the Gallager expurgated bound, there is also a passage from a curvy part at low rates to a straight–line part at high rates. However, in Gallager's derivation, the passage happens due to a more technical reason. Since Gallager's analysis is based on the inequality $[\sum_{m'} a_{m'}]^{1/\rho} \leq \sum_{m'} a_{m'}^{1/\rho}$, which holds only for $\rho \geq 1$, the maximization over ρ is a–priori limited to the range $\rho \geq 1$. The linear part of the curve is then generated due to the fact that for higher rates, the unconstrained achiever of $E_{ex}(R)$ is $\rho^* < 1$, and so, the constrained one remains $\rho^* = 1$, independently of R in this range.

References

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

[2] I. Csiszár and J. Körner, *Information Theory – Coding Theorems for Dsicrete Memoryless Systems*, Academic Press, New York, 1981.

[3] I. Csiszár, J. Körner, and K. Marton, "A new look at the error exponent of a discrete memoryless channel," *Proc. ISIT '77*, p. 107 (abstract), Cornell University, Itacha, New York, U.S.A., 1977.

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

[5] B. Derrida, "The random energy model," *Physics Reports* (Review Section of Physics Letters), vol. 67, no. 1, pp. 29–35, 1980.

[6] B. Derrida, "Random–energy model: an exactly solvable model for disordered systems," *Phys. Rev. B*, vol. 24, no. 5, pp. 2613–2626, September 1981.

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

[8] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. on Inform. Theory*, vol. IT–11, pp. 3–18, January 1965.

[9] R. G. Gallager, *Information Theory and Reliable Communication*, New York, Wiley 1968.

[10] Y. Kaspi and N. Merhav, "Error exponents for broadcast channels with degraded message sets," *IEEE Trans. Inform. Theory*, vol. 57, no. 1, pp. 101–123, January 2011.

[11] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1261–1269, July 1993.

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

[13] N. Merhav, "Relations between random coding exponents and the statistical physics of random codes," *IEEE Trans. Inform. Theory*, vol. 55, no. 1, pp. 83–92, January 2009.

[14] N. Merhav, "The generalized random energy model and its application to the statistical physics of ensembles of hierarchical codes," *IEEE Trans. Inform. Theory*, vol. 55, no. 3, pp. 1250–1268, March 2009.

[15] N. Merhav, "Statistical physics and information theory," (invited paper) *Foundations and Trends in Communications and Information Theory*, vol. 6, nos. 1–2, pp. 1–212, 2009.

[16] N. Merhav, "On the statistical physics of directed polymers in a random medium and their relation to tree codes," *IEEE Trans. Inform. Theory*, vol. 56, no. 3, pp. 1345–1350, March 2010.

[17] N. Merhav, "Rate–distortion function via minimum mean square error estimation," *IEEE Trans. Inform. Theory*, vol. 57, no. 6, pp. 3196–3206, June 2011.

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

[19] J. K. Omura, "Expurgated bounds, Bhattacharyya distance, and rate distortion functions," *Information and Control*, vol. 24, pp. 358–383, 1974.

[20] C. E. Shannon, R. G. Gallager and E. R. Berlekamp, " Lower bounds to error probability for coding on discrete memoryless channels. I" *Information and Control*, vol. 10, pp. 65–103, January 1967.

[21] C. E. Shannon, R. G. Gallager and E. R. Berlekamp, " Lower bounds to error probability for coding on discrete memoryless channels. II" *Information and Control*, vol. 10, pp. 522–552, May 1967.

[22] E. Sabbag and N. Merhav, "Error exponents of optimum erasure/list and ordinary decoding for channels with side information," *Proc. ISIT 2012*, pp. 2949–2953, Cambridge, MA, U.S.A., July 2012.

[23] A. Somekh–Baruch and N. Merhav, "Exact random coding error exponents for erasure decoding," *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 6444–6454, October 2011.

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