Multi-Class Source-Channel Coding

Irina E. Bocharova, Senior Member, IEEE, Albert Guillén i Fàbregas, Senior Member, IEEE,
Boris D. Kudryashov, Senior Member, IEEE, Alfonso Martinez, Senior Member, IEEE,
Adrià Tauste Campo, Member, IEEE, Gonzalo Vazquez-Vilar, Member, IEEE

Abstract

We study a source-channel coding scheme in which source messages are assigned to different classes and encoded using a channel code that depends on the class index. The performance of this scheme is studied by means of random-coding error exponents and validated by simulation of a low-complexity implementation using existing source and channel codes. While each class code can be seen as a concatenation of a source code and a channel code, the overall performance improves on that of separate source-channel coding and approaches that of joint source-channel coding when the number of classes increases.

I. INTRODUCTION

Implicit in Shannon’s source-channel coding theorem [1] is the fact that reliable transmission of a source through a channel can be accomplished by using separate source and channel codes. This means that a concatenation of a (channel-independent) source code followed by a (source-independent) channel code achieves vanishing error probability as the block length goes to infinity, as long as the source entropy is smaller than the channel capacity [1]. However, in the non-asymptotic regime an optimally designed joint source-channel code can perform strictly better than a separate code. This improvement (i.e., reduction in error probability) has been quantified in terms of error exponents [2], [3] and in terms of source and channel dispersion [4], [5]. Joint design has an error exponent at most twice of that of separate codes [6], and a dispersion gain that depends on the target error probability; for

I. E. Bocharova and B. D. Kudryashov are with the St. Petersburg University of Information Technologies, Mechanics and Optics, St. Petersburg, Russia (emails: {irina, boris}@eit.lth.se). A. Guillén i Fàbregas is with the Institució Catalana de Recerca i Estudis Avançats (ICREA), the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, and the Department of Engineering, University of Cambridge, Cambridge, United Kingdom (email: guillen@ieee.org). A. Martinez, A. Tauste Campo, are with the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain (emails: alfonso.martinez@ieee.org, adria.tauste@upf.edu). G. Vazquez-Vilar was with the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain. Now he is with the Signal Theory and Communications Department, Universidad Carlos III de Madrid, Leganés, Spain (email: gvazquez@ieee.org).

This work has been funded in part by the European Research Council under ERC grant agreement 259663, by the European Union’s 7th Framework Programme under grant agreements 303633 and 329837, and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150 and TEC2012-38800-C03-03.

This work was presented in part at the 2014 IEEE Symposium on Information Theory, Honolulu, HI, June 29–July 4, 2014, and at the 8th International Symposium on Turbo Codes and Iterative Information Processing, Bremen, Germany, Aug. 18–22, 2014.
vanishing values of the latter, the dispersion of joint design is at best half of the dispersion of separate design [5]. This potential gain justifies the interest in practical finite-length joint source-channel codes.

Several practical joint source-channel coding schemes have been proposed in the past. One possible approach is to adapt existing channel coding techniques to exploit the knowledge on the source statistics at the decoder side. For example, [7] proposes a modification of the Viterbi decoding algorithm to use the a priori probabilities of the source bits, [8] uses punctured turbo-codes with a modified iterative decoder, and [9] uses both source and channel LDPC codes with a decoder exploiting the joint graph structure of the codes and the source. Other schemes exploit the source statistics both at the encoder and decoder. In [10], source bits are matched to a non-systematic LDPC code via scrambling or splitting. In [11]–[13] the authors propose a trellis-structure description of the Huffman code and an appropriate channel code so that joint decoding is possible. This technique has been extended to arithmetic and Lempel-Ziv source coding in [14] and [15], respectively. A different technique based on unequal error protection has been proposed by Shkel et al. in [16]. Common to these source-channel coding schemes is the underlying idea of approximating the (optimum) maximum a posteriori (MAP) decoder by using certain properties of the source statistics.

In this paper, we analyze a novel source-channel coding scheme in which source messages are assigned to disjoint classes and encoded by codes that depend on the class index. Under MAP decoding, this scheme attains the joint source-channel reliability function in the cases where it is known to be tight [17]. In this work, however, we are interested in characterizing the performance of this coding scheme under simpler, sub-optimal decoding. First, we process the channel output in parallel for each class using a bank of maximum likelihood (ML) decoders. Then, the decoded message is selected from the outputs of the ML decoders based on a MAP criterion. While this construction fails to achieve the best performance of joint source-channel coding, for a fixed number of classes, it can be implemented with reduced complexity using existing source and channel codes. This scheme is shown to improve on the error exponent of separate coding, and, as the number of classes increases, to approach the error exponent of joint source-channel coding [3].

In our scheme, messages are assigned to classes based on their probability. The most probable messages are encoded with a low-rate channel code, and hence they receive an increased protection against channel errors. Analogously, less probable messages are assigned to classes that receive less protection against channel errors. If we could apply optimum unequal error protection in a message-by-message basis (i.e., emulating a MAP decoder) we would be able to achieve the best joint source-channel coding performance [17]. However, due to complexity constraints, we assign source messages of different probabilities to the same class and, hence, we only approximate the best joint scheme. In contrast with [16], where unequal error protection is guaranteed via the code construction, in this work we perform unequal error protection at the decoding stage only. While we restrict our analysis to memoryless sources and channels, the proposed construction can be applied to more general cases where the source and/or the channel have memory.

The structure of the paper is as follows. In Section II we present the system model and our multi-class source channel coding scheme. Section III presents a random-coding analysis of this scheme. Section IV validates these
results by means of simulation of a reduced complexity implementation based on LDPC codes. Finally, we present some concluding remarks in Section V.

II. SYSTEM MODEL AND NOTATION

We consider the transmission of a length-$k$ discrete memoryless source over a memoryless channel using length-$n$ block codes. The source output is distributed according to $P^k(v) = \prod_{i=1}^{k} P(v_i)$, $v = (v_1, \ldots, v_k) \in \mathcal{V}^k$, where $P(v)$ is the source symbol distribution, and $\mathcal{V}$ is a discrete alphabet. Without loss of generality, we assume that $P^k(v) > 0$ for all $v$.

The channel law is given by $W^n(y|x) = \prod_{i=1}^{n} W(y_i|x_i)$, $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, $y = (y_1, \ldots, y_n) \in \mathcal{Y}^n$, where $W(y|x)$ denotes the channel transition probability and $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output alphabet, respectively. We define $t \triangleq \frac{k}{n}$.

A source-channel code is defined by an encoder and a decoder. The encoder maps the message $v$ to a length-$n$ codeword $x(v)$. Based on the channel output $y$, the decoder selects a message $\hat{v}(y)$. Throughout the paper, random variables will be denoted by capital letters and the specific values they take are denoted by the corresponding lower case letters. When clear from context, we avoid writing the dependence of the decoder output on the channel output explicitly. The error probability of a source-channel code is thus given by

$$\epsilon_n = \Pr\{V \neq \hat{V}\}. \quad (1)$$

We characterize this probability in terms of error exponents. An exponent $E > 0$ is to said to be achievable if there exists a sequence of codes with $n = 1, 2, \ldots$, and $k = 1, 2, \ldots$, whose error probabilities $\epsilon_n$ satisfy

$$\epsilon_n \leq e^{-nE+o(n)}, \quad (2)$$

where $o(n)$ is a sequence such that $\lim_{n \to \infty} \frac{o(n)}{n} = 0$. The supremum of all achievable exponents is usually referred to as reliability function.

If the encoder (resp. the decoder) is the concatenation of a source and channel encoder (resp. channel and source decoder), we recover separate source-channel coding. In this work, we propose a new scheme in which the source-message set is split into subsets, and concatenated source and channel codes are used for each subset. At the receiver, each channel code is decoded in parallel, and the final output is selected based on MAP criterion. A block diagram of the scheme is shown in Fig. 1.

For each $k$, we define a partition $\mathcal{P}_k$ of the source-message set $\mathcal{V}^k$ into $N_k + 1$ disjoint subsets $\mathcal{A}_i^k$, $i = 0, 1, \ldots, N_k$. We shall refer to these subsets as classes. Sometimes, we consider sequences of sources, channels and partitions.

Figure 1. Block diagram of the proposed multi-class source-channel coding scheme.
where $N_k$ grows with $k$. In this case, we assume that this variation is subexponential in $k$. The asymptotic number of classes as $k \to \infty$ is $N \triangleq \lim_{k \to \infty} N_k$, hence $N \in \mathbb{N} \cup \{\infty\}$. More specifically, we consider a sequence of partitions such that, for each $k$, source messages are assigned to classes depending on their probability,

$$A^k_i = \{ v \mid \gamma^k_i < P^k(v) \leq \gamma^k_{i+1} \}, \quad i = 0, \ldots, N_k,$$

with $0 = \gamma^k_0 \leq \gamma^k_1 \leq \ldots \leq \gamma^k_{N_k+1} = 1$. The thresholds $\gamma_1, \ldots, \gamma_{N_k}$ should be properly selected to optimize the system performance. If $P^k(v)$ were equal to 0 for some $v$, we could always define a new source without these messages and use the partition (3).

All the messages belonging to the class $A^k_0$ are encoded with the same codeword $x(v) = x_0$ and are assumed to lead to a decoding error. For each remaining class $A^k_i$, messages are encoded with a channel code $C_i$ of rate $R_i \triangleq \frac{1}{n} \log |A^k_i|, \quad i = 1, \ldots, N_k$.

Upon receiving a message $v$, the encoder transmits the codeword $x(v)$. At the receiver, we use a two-step decoder (see Fig. 1). For each class $A^k_i$, $i = 1, \ldots, N_k$, the $i$-th ML decoder selects a message $\hat{v}_i$ in $A^k_i$ as

$$\hat{v}_i = \arg \max_{v \in A^k_i} W^n(y|x(v)).$$

Next, the decoder selects from the set $\{ \hat{v}_i \}_{i=1}^{N_k}$, the source message with largest MAP decoding metric

$$q(v, y) \triangleq P^k(v)W^n(y|x(v)).$$

That is, the final output is $\hat{v} = \hat{v}_\hat{i}$, where the class index selected by the MAP decoder corresponds to

$$\hat{i} = \arg \max_{i = 1, \ldots, N_k} q(\hat{v}_i, y).$$

In the following, we find useful to define the channel coding and source coding exponents. Let the Gallager’s channel and source functions be given by

$$E_0(\rho, Q) \triangleq -\log \sum_y \left( \sum_x Q(x)W(y|x)^{1+\rho} \right)^{1+\rho},$$

and

$$E_s(\rho) \triangleq \log \left( \sum_v P(v)^{1+\rho} \right)^{1+\rho},$$

respectively. For channel coding alone, the random-coding exponent at rate $R$ for an input distribution $Q$ is achievable and it is given by [2]

$$E_i(R, Q) = \max_{\rho \in [0,1]} \left\{ E_0(\rho, Q) - \rho R \right\}.$$  

We define $E_i(R) \triangleq \max_Q E_i(R, Q)$. For source coding alone, the reliability function of a source $P$ at rate $R$, denoted by $e(R)$, is given by [18]

$$e(R) = \sup_{\rho \geq 0} \{ \rho R - E_s(\rho) \}.$$
III. ERROR EXPONENT ANALYSIS

The two-step decoding scheme in Section II includes both separate and joint source-channel coding as special cases. For $N_k = 1$, the scheme corresponds to a concatenation of an almost-lossless source code and a channel code $C_1$ with intermediate rate $R_1 = \frac{1}{n} \log |A_i^k|$. Similarly, when $A_i^k$ is empty and the subsets $A_i^k$, $i = 1, \ldots, N_k$, coincide with the source type classes for all $k$, the overall decoder is actually MAP, and the proposed scheme attains the best known achievable error exponent [3], [19]. In this section, we analyze the error exponent of the proposed scheme for an arbitrary number of classes. We start by describing a number of properties of our partition of the source message set.

For a sequence of partitions $\{A_i^k\}$, $k = 1, 2, \ldots$, we define the following functions

$$E_{s,i}^k(\rho) \triangleq \log \left( \sum_{v \in A_i^k} P^k(v)^{1+\rho} \right)^{1+\rho},$$

$$E_{s,i}(\rho) \triangleq \lim_{k \to \infty} \frac{1}{k} E_{s,i}^k(\rho).$$

These functions take over the role of Gallager’s source function $E_s(\cdot)$ when dealing with multiple classes (see, e. g., [17]). In principle, the functions $E_{s,i}(\cdot)$ are difficult to evaluate, since they involve the computation of a limit. The following result provides a simple characterization of $E_{s,i}(\cdot)$ for a sequence of partitions of the form (3). For $\rho \geq 0$, we denote the derivative of $E_s(\rho)$ evaluated at $\rho$ as

$$E_s'(\rho) \triangleq \frac{\partial E_s(\rho)}{\partial \rho} \bigg|_{\rho = \rho}.$$

and we define the tilted distribution

$$P_\rho(v) \triangleq \frac{P(v)^{1+\rho}}{\sum_{v} P(v)^{1+\rho}}.$$

**Lemma 1:** For a discrete memoryless source $P$ and the partition given in (3), it holds that

$$E_{s,i}(\rho) = \begin{cases} E_s(\rho_{i+1}^*) + (\rho - \rho_{i+1}^*) E_s'(\rho_{i+1}^*), & \rho < \rho_{i+1}^*, \\ E_s(\rho), & \rho_{i+1}^* \leq \rho \leq \rho_i^*, \\ E_s(\rho_i^*) + (\rho - \rho_i^*) E_s'(\rho_i^*), & \rho > \rho_i^*, \end{cases}$$

for $i = 0, \ldots, N$, where $\rho_i^*$ is given by the solution to the implicit equation

$$\sum_{v} P_{\rho_i^*}(v) \log P(v) = \log \gamma_i,$$

as long as $\sum_{v} \frac{1}{|V|} \log P(v) < \log \gamma_i < \sum_{v} P(v) \log P(v)$. In case that $\log \gamma_i \leq \sum_{v} \frac{1}{|V|} \log P(v)$ then $\rho_i^* = \infty$; if $\log \gamma_i \geq \sum_{v} P(v) \log P(v)$, then $\rho_i^* = 0$.

**Proof:** See Appendix I.

An example of the characterization in Lemma 1 is shown in Fig. 2 for a three-class partition. We observe that $E_{s,i}(\rho)$ is equal to $E_s(\rho)$ for the interval $\rho_{i+1}^* \leq \rho \leq \rho_i^*$, and corresponds to a straight line tangent to $E_s(\rho)$ out of those intervals. Since the thresholds $\gamma_0$ and $\gamma_{N+1}$ are fixed to 0 and 1, respectively, then $\rho_0^* = \infty$ and $\rho_{N+1}^* = 0$. 

November 3, 2014

DRAFT
For the remaining thresholds, given the continuity of the left- and right-hand-side of (17), we can obtain any finite value of $\rho_i^* \in [0, \infty)$ by appropriately choosing the threshold $\gamma_i \in [0, 1], i = 1, \ldots, N$.

The slope of the function $E_{s,i}(\rho)$ as $\rho \to \infty$ is such that (see Appendix II for details)

$$\lim_{\rho \to \infty} \frac{1}{\rho} E_{s,i}(\rho) = E'_s(\rho^*_i)$$

$$= R_i$$

Equations (18)-(19) describe a one-to-one relation between $\rho^*_i$, and the asymptotic rate $R_i$ of the $i$-th class. Hence, it is asymptotically equivalent to optimize the partition over any of these three sets of parameters: $\{\gamma_i\}, \{\rho^*_i\}$ or rates $\{R_i\}$.

Using (12) and (13), we can represent the asymptotic probability of the set $A^c_i$ by means of the exponent

$$\lim_{k \to \infty} \frac{1}{k} \log \left( \sum_{v \in A^c_i} P_k(v) \right) = E_{s,i}(0)$$

$$= E_s(\rho^*_{i+1}) - \rho^*_{i+1} E'_s(\rho^*_{i+1})$$

$$= \max_{\rho \in [0, \infty)} \left\{ E_s(\rho) - \rho \frac{R_{i+1}}{t} \right\}$$

$$= -e \left( \frac{R_{i+1}}{t} \right)$$

where (21) follows from Lemma 1, in (22) we used the identity in (19) and the fact that $\rho^*_{i+1}$ is the point where $E_s(\rho)$ has slope $\frac{R_{i+1}}{t}$, i.e., it maximizes the quantity in brackets, and in (23) the definition (11) of the error exponent of a discrete source memoryless source compressed to rate $E'_s(\rho^*_{i+1}) = \frac{R_{i+1}}{t}$. This expression confirms the intuition that lower (resp. higher) rates are used to encode messages with high (resp. low) probabilities.
When the classes correspond to source-type classes [20], we obtain
\[
E_{k,i}(\rho) = -D(P_i \parallel P) + \rho H(P_i),
\]
where \(P_i\) denotes the empirical distribution of the corresponding source-type class. Since any distribution \(P\) can be written as the limit of a sequence of types [20, Sec. IV], it follows that for source-type classes \(\{E_{k,i}(\rho)\}\) corresponds to the set of straight lines tangent to the original \(E_s(\rho)\) function. In particular, we have that \(E_{k,i}(0) = -D(P_i \parallel P)\) and \(\lim_{\rho \to \infty} \frac{1}{\rho} E_{k,i}(\rho) = H(P_i)\), as they should for types [20, Lemma II.2].

We now proceed to derive a lower bound to the error exponent of the multi-class scheme. To this end, we define three different error events. The first occurs when a source message belongs to the dummy set \(\mathcal{E}_s\) (source error); the second occurs when, for a source message belonging to class \(\mathcal{A}_i^k\), the \(i\)-th ML decoder makes an error (ML error); and the third occurs when the \(i\)-th ML decoder output is correct but the MAP decoder makes an error (MAP error). More precisely, these three error events are defined respectively as
\[
\mathcal{E}_s \triangleq \{v \in \mathcal{A}_0^k\},
\]
\[
\mathcal{E}_{\text{ML}}(i) \triangleq \{v \in \mathcal{A}_i^k, \hat{v}_i \neq v\},
\]
\[
\mathcal{E}_{\text{MAP}}(i) \triangleq \{v \in \mathcal{A}_i^k, \hat{v}_i = v, \hat{i} \neq i\}.
\]

Then, using that these error events are disjoint, the error probability can be written as
\[
\epsilon_n = \Pr \left\{ \mathcal{E}_s \cup \left( \bigcup_{i=1}^{N_k} \mathcal{E}_{\text{ML}}(i) \right) \cup \left( \bigcup_{i=1}^{N_k} \mathcal{E}_{\text{MAP}}(i) \right) \right\}
= \Pr \{ V \in \mathcal{A}_0^k \} + \sum_{i=1}^{N_k} \Pr \{ V \in \mathcal{A}_i^k, \hat{V}_i \neq V \} + \sum_{i=1}^{N_k} \Pr \{ V \in \mathcal{A}_i^k, \hat{V}_i = V, \hat{i} \neq i \},
\]
(29)

To obtain lower bounds on the error exponent of this construction, we start by upper-bounding every term in the third summand in (29) as
\[
\Pr \{ V \in \mathcal{A}_i^k, \hat{V}_i = V, \hat{i} \neq i \} = \Pr \{ V \in \mathcal{A}_i^k, \hat{i} \neq i | \hat{V}_i = V \} \Pr \{ \hat{V}_i = V \}
\leq \Pr \{ V \in \mathcal{A}_i^k, \hat{i} \neq i | \hat{V}_i = V \} \Pr \{ V \}
\leq \Pr \{ V \in \mathcal{A}_i^k, q(V, Y) \leq \max_{\hat{v} \neq V, \hat{v} \neq \mathcal{A}_0^k} q(\hat{v}, Y) \},
\]
(32)

where (30) follows from the chain rule, (31) from upper bounding \(\Pr \{ \hat{V}_i = V \} \) by \(\Pr \{ V \} \), and (32) follows from (7) by enlarging the set of source messages over which the maximum is computed, assuming that ties are decoded as errors, and using the chain rule again.

Substituting (32) into (29), we obtain
\[
\epsilon_n \leq \Pr \{ V \in \mathcal{A}_0^k \} + \sum_{i=1}^{N_k} \Pr \{ V \in \mathcal{A}_i^k, \hat{V}_i \neq V \} + \Pr \{ V \notin \mathcal{A}_0^k, q(V, Y) \leq \max_{\hat{v} \neq V, \hat{v} \neq \mathcal{A}_0^k} q(\hat{v}, Y) \}.
\]
(33)

For each \(k, n\), assign the distribution \(Q_i(x)\) to each class, \(\mathcal{A}_i^k\), \(i = 0, \ldots, N_k\). Then, for each source message \(v \in \mathcal{A}_i^k\), we may randomly generate the codeword \(x(v)\) according to \(Q_i^v(x) \triangleq \prod_{j=1}^{N_k} Q_i(x_j), i = 1, \ldots, N_k\). For the
class $A_0^k$ whose messages we discard, $Q_0$ denotes a symbol distribution that assigns mass 1 to a predetermined null symbol. Then, its Gallager’s function satisfies that $E_0(\rho_0, Q_0) = 0$ for any $\rho_0 \in [0, 1]$. We also define $R_{N+1} \triangleq 0$ such that $e \left( \frac{R_{N+1}}{t} \right) = 0$. The next result follows from (33) using the exponential bounds [18, Th. 5.2], [2, Th. 5.6.1] and [17, Th. 1] via the random-coding argument.

**Theorem 1:** There exists a sequence of codes, partitions and decoders as defined in Section II that achieve the following exponent

$$\min_{i=0, \ldots, N} \left\{ E_r(R_i, Q_i) + te \left( \frac{R_{i+1}}{t} \right) \right\},$$

(34)

where the rates $R_i$, $i = 0, \ldots, N$, are related to the partition parameters in Lemma 1 and $N = \lim_{k \to \infty} N_k$.

**Proof:** See Appendix III.

The achievable error exponent in Theorem 1 corresponds to that in [21, Th. 1]. Under certain assumptions it coincides with an upper bound to the error exponent of this code construction [21, Th. 2]. This is the case for a given class of channels (such as the binary symmetric channel, binary erasure channel or phase-shift-keying modulated additive white Gaussian noise channel (AWGN)), when the intermediate rates optimizing (34) are above the critical rate of the channel and the codes $C_1, \ldots, C_{N_k}$ are linear. While this converse result only applies to a class of codes and channels, it shows that in these cases there is no loss in exponent by considering the bound in Theorem 1.

Further analysis involves optimization over rates $R_i$ (i.e., thresholds $\gamma_i$) and distributions $Q_i$, $i = 1, \ldots, N$. The bound in Theorem 1 can be relaxed to obtain an alternative expression. We define $E_0(\rho) \triangleq \max_Q E_0(\rho, Q)$.

**Theorem 2:** There exists a sequence of codes, partitions and decoders defined in Section II with $N \geq 2$ that achieve the following exponent,

$$\max_{R \geq R' \geq 0} \min \left\{ \max_{\rho \geq 0} \left\{ \rho R - tE_s(\rho) \right\}, \right.$$ 

$$\max_{\rho' \in [0, 1]} \left\{ E_0(\rho') - tE_s(\rho') - \rho' \frac{R - R'}{N - 1} \right\},$$

$$\max_{\rho'' \in [0, 1]} \left\{ E_0(\rho'') - \rho'' R' \right\},$$

(35)

Moreover, the rate of each class in the partition is

$$R_i = R' + (i - 1) \frac{R - R'}{N - 1}, \quad i = 1, \ldots, N,$$

(36)

where $R$ and $R'$ are the values optimizing (35).

**Proof:** See Appendix IV.

The bound in Theorem 2 is simple to evaluate since it only involves the well known functions $E_s(\cdot)$ and $E_0(\cdot)$, and the optimization is performed over a fixed number of parameters $(\rho, \rho', \rho'', R$ and $R')$ independent of $N$. Furthermore, as we show next with an example, it is indistinguishable of the bound from Theorem 1 in several cases of interest.
For \( N = 1 \), it follows that \( E_r(R_0, Q_0) = 0 \) and \( e(R_2) = 0 \). Optimizing (34) over intermediate rate \( R = R_1 \) and distribution \( Q_1 \), Theorem 1 recovers the separate source-channel exponent [3],

\[
\max_{R \geq 0} \min_{t \in [0,1]} \left\{ E_r(R), tE_s(R) \right\}.
\]

(37)

Let \( N_k \) grow (subexponentially) with \( k \) in such a way that \( \lim_{k \to \infty} N_k = \infty \). For this discussion only, we allow \( R \) and \( R' \) to depend on \( k \) as \( R_k \) and \( R'_k \), respectively. Let us choose the sequences \( R_k \) and \( R'_k \) such that \( \lim_{k \to \infty} R_k = \infty \), \( \lim_{k \to \infty} R'_k = 0 \) and \( \lim_{k \to \infty} \frac{R_k - R'_k}{N_k - 1} = 0 \), i.e., \( R_k = o(N_k) \). In this case, the first and last terms within the minimization in (35) become irrelevant and the bound in Theorem 2 recovers Gallager’s source-channel error exponent [2, p. 534, Prob. 5.16],

\[
\max_{\rho \in [0,1]} \{ E_0(\rho) - t E_s(\rho) \}.
\]

(38)

In several cases of interest, the exponent (38) coincides with the joint source-channel reliability function. However, for specific source and channel pairs the following exponent \([3], [6]\) gives a tighter bound to the reliability function,\n
\[
\min_{R \geq 0} \left\{ E_r(R) + tE_s(R) \right\} = \max_{\rho \in [0,1]} \{ E_0(\rho) - t E_s(\rho) \},
\]

(39)

where \( E_0(\rho) \) denotes the concave hull of \( E_0(\rho) \), defined pointwise as the supremum over convex combinations of any two values of the function \( E_0(\rho) \) [22, p. 36]. While the bound in Theorem 2 does not attain (39), this error exponent can be recovered from Theorem 1 by identifying the classes with the source-type classes. In this case, \( R_i = tH(\mathbf{P}_i) \) and \( R_{i+1} = tH(\mathbf{P}_{i+1}) \) become infinitely close to each other and they uniformly cover the interval \( [0, t \log(|V|)] \) for \( i = 1, 2, \ldots \). As a result, (34) recovers the left-hand side of (39). This shows that the gap between the bounds in Theorems 1 and 2 can be strictly positive.

A. Example

A binary memoryless source (BMS) with parameter \( p \triangleq P(1) \leq 1/2 \) is to be transmitted over a binary-input AWGN channel with signal-to-noise ratio (SNR) \( E_b/N_0 \). We normalize \( E_b/N_0 \) with respect to the number of transmitted information bits, i.e. \( k/n \) times the source entropy \( H(V) \). Let \( h_2(p) = -p \log_2 p - (1-p) \log_2(1-p) \) denote the binary entropy function in bits. We define a signal-to-noise ratio per source bit \( E_b/N_0 \) as

\[
\frac{E_b}{N_0} \triangleq \frac{n}{kh_2(p) N_0}.
\]

(40)

Figure 3 shows the achievable error exponents for different coding schemes as a function of \( E_b/N_0 \) in decibels. The error exponents in the figure correspond to separate source-channel coding (37), joint source-channel coding (38), and the proposed multi-class scheme with \( N = 2, 3, 5, 12 \). The bound in Theorem 1 has been optimized over the parameters \( \rho_i, i = 0, \ldots, N \), and thresholds \( \gamma_i, i = 1, \ldots, N \). The bound in Theorem 2 has been optimized over the parameters \( \rho, \rho', \rho'', R \) and \( R' \). In both cases the channel input distribution has been chosen to be equiprobable. From the figure we can see that the bound in Theorem 1 and the relaxed version in Theorem 2 coincide for \( N = 2, 3 \). For \( N = 2 \), the multi-class scheme shows a 0.4-0.7 dB improvement over separate source coding, with just a small increase in complexity. Moreover, from the curves for \( N = 2, 3, 5, 12 \) we can see that the multi-class
construction approaches the joint source-channel error exponent as the number of classes increases, confirming the result of Theorem 2 since (38) and (39) coincide for this example.

IV. PRACTICAL CODE DESIGN

In this section, we illustrate how the proposed construction can be used to design joint source-channel codes. We consider the transmission of a BMS with $P(1) \leq 1/2$ over a binary-input AWGN channel. The proposed scheme is a two-class code composed of a fixed-to-variable lossless source code followed by two linear $(n, k_i)$-codes $C_i$, $i = 1, 2$. The lossless source code corresponds to the class selector in Fig. 1. The ML decoders in Fig. 1 can be replaced by using standard decoding techniques. This fixed-to-variable-to-fixed source-channel code allows a simple implementation of the proposed scheme using existing source and channel codes.

As a source code we use a fixed-to-variable coding scheme that assigns shorter codewords to the most probable messages, i.e., messages with smallest Hamming weight. Two examples are enumerative [23] and arithmetic coding [24]. For a source message $v$, the length of source codeword $L(v)$ determines which one of two available codes will be used to encode each source message. Note that since the source code is assumed lossless, this is equivalent to assign source messages to classes based on their probability. If $L(v) \leq k_1$ the channel code $C_1$ is used for transmission, otherwise, if $L(v) \leq k_2$ the second code, $C_2$, is used. If $L(v) > k_2$ a source coding error is reported. In this coding scheme, $k_i - L(v)$ leftover bits may appear due to a mismatch between the source and channel code, $i = 1, 2$. These bits can be used to include an additional redundancy check (see [25]), however, for simplicity, here we assume that they are padded with zeros. Due to these leftover bits, we do not use all the codewords belonging to each of the channel codes, in contrast to the theoretical analysis from Section III. However, in general, $k_i \approx \log_2|A_i^k|$, $i = 1, 2$, and the performance loss is negligible.

At the decoder, two decoding attempts are performed in parallel. In each branch, corresponding to one of the two possible codes, the receiver performs ML (or quasi-ML) decoding. Both decoder outputs are then checked.
to verify whether they are valid source sequences. If only one of the two outputs is found to be compatible, the corresponding data are used as the message. If both decoders fail, a predetermined message, for example the all-zero data sequence, is used. Finally, if both source decoders report success, the message with larger a posteriori likelihood is selected.

A. Code Optimization

The specific pair of $(n, k)$-codes depends on the signal-to-noise ratio $E_b/N_0$. Obviously, the choice of the code rates and of the codes themselves is critical for the system performance. If the block length $n$ is small we can obtain a set of good channel codes with different coding rates using techniques from, e. g. [26]–[28]. Then, for each $E_b/N_0$ the best pair of codes from this set can be selected by simulating the system performance. While this optimization procedure is feasible for short block lengths, it can become computationally intractable as the block length or rate granularity grow large.

In these cases, we may resort to the exponents from Section III to estimate the optimal coding rate pair. To this end we compute the rates $R_1$ and $R_2$ optimizing either Theorem 1 or Theorem 2, and select two codes of rates $R_1$ and $R_2$. Since the exponential behavior dominates for large block lengths, these rates become asymptotically optimal as the block length grows large. As we will see in the simulations section, Theorems 1 and 2 give a good approximation of the optimal coding rates for moderate block lengths ($n \approx 1000$).

B. Lower bound on the error probability

We now derive a lower bound on the error probability of a two-class linear coding scheme for a BMS. This lower bound will serve as a benchmark to the performance of the optimize two-class code described above.

Disregarding the last summand in (29) we lower-bound the error probability of a given code as

$$
\epsilon_n \geq \Pr\{\mathbf{V} \in \mathcal{A}_0^k\} + \sum_{i=1,2} \Pr\{\mathbf{V} \in \mathcal{A}_i^k, \hat{\mathbf{V}}_i \neq \mathbf{V}\} + \sum_{i=1,2} \Pr\{\mathbf{V} \in \mathcal{A}_i^k\} \Pr\{\hat{\mathbf{V}}_i \neq \mathbf{V} | \mathbf{V} \in \mathcal{A}_i^k\}.
$$

A lower bound on the error probability of a channel code of rate $R$ is given by Shannon’s sphere-packing bound [29]. Let codewords be distributed over the surface of an $n$-dimensional hypersphere with squared radius $E = nE_s$ and centered at the origin of coordinates. Let $\theta$ be the half-angle of a cone with vertex at the origin and with axis going through one arbitrary codeword. We let $Q(\theta)$ denote the probability that such codeword be moved outside the cone by effect of the Gaussian noise. We choose $\theta_{n,R}$ such that the solid angle subtended by a cone of half-angle $\theta_{n,R}$ is equal to $\Omega_n/2^{nR}$, where $\Omega_n$ is the surface of the $n$-dimensional hypersphere. Then, $Q(\theta_{n,R})$ is a lower bound on the error probability of a channel code of length $n$ and rate $R$. In particular, in the proposed multiclass scheme, this expression provides a lower bound on the error probability of each of the linear codes under ML decoding (when ties are resolved randomly), i. e.,

$$
\Pr\{\hat{\mathbf{V}}_i \neq \mathbf{V} | \mathbf{V} \in \mathcal{A}_i^k\} \geq Q(\theta_{n,R_i}), \ i = 1, 2.
$$
This bound is accurate for low SNRs and relatively short codes [30]. In order to compute (43) we use the approximation from [31], known to be accurate for error probabilities below 0.1.

For a BMS with \( p = P(1) \leq 1/2 \), it is possible to obtain a closed-form expression for the source terms \( \Pr\{ V \in A^k_i \} \), \( i = 0, 1, 2 \). Consider a class \( A^k_i \) composed by the sequences with Hamming weights \( w \in [w_1, w_2] \), where \( w_1 \) and \( w_2 \) are two arbitrary integers. Then, it follows that

\[
\Pr\{ V \in A^k_i \} = B_{k,p}(w_1, w_2),
\]

(44)

where we defined

\[
B_{k,p}(w_1, w_2) \triangleq \sum_{w=w_1}^{w_2} \binom{k}{w} p^w (1-p)^{k-w}.
\]

(45)

The best coding strategy is to encode the sequences of Hamming weight \( w \in \{0, \ldots, w_1\} \) with the first (lower-rate) channel code and the sequences of weight \( w \in \{w_1 + 1, \ldots, w_2\} \) with the second (higher-rate) code. All other sequences are transmitted by some fixed codeword which leads to decoding error. Therefore, using (43) and (44) in (42), we obtain

\[
\epsilon_n \geq \min_{\substack{w_1=0, \ldots, k, \\
w_2=w_1+1, \ldots, k}} \left\{ B_{k,p}(0, w_1)Q(\theta_n, R(0, w_1)) + B_{k,p}(w_1 + 1, w_2)Q(\theta_n, R(w_1 + 1, w_2)) + B_{k,p}(w_2 + 1, k) \right\},
\]

(46)

where the rate \( R(w_1, w_2) \) is given by

\[
R(w_1, w_2) = \frac{1}{n} \left\lfloor \log_2 \sum_{w=w_1}^{w_2} \binom{k}{w} \right\rfloor.
\]

(47)

C. Simulation Results

In this subsection we show simulation results for different implementations of a two-class scheme in short and moderate block length scenarios. The source probability is fixed to \( P(1) = 0.1 \).

1) Short block length scenario \((k = 80, n \approx 100)\):

Figure 4 shows the simulated FER performance of an implementation using tail-biting codes and ML decoding. As source code we use an enumerative coding scheme and as channel codes we have chosen a family of tail-biting (TB) codes of rates \( R = 1/2, 3/5 \) and \( 3/4 \). The code of rate \( R = 1/2 \) was taken from [32], and the codes of rates \( 3/5 \) and \( 3/4 \) where chosen by doing a short search for high-rate convolutional codes using techniques from [26], [27]. Among the most efficient ML decoding algorithms we have selected BEAST [33] which allows ML decoding for codes of length 100 with acceptable complexity. The curves “Separate” and “Two-class code” show the best performance obtained within the corresponding family of codes. The two-class scheme outperforms separate coding by about 1 dB, in agreement with the values predicted by the random coding analysis. Also, from the figure we see that the lower bound (46) can be used to predict not only the gain value but also the best error probability.

Table I shows the best code rate pairs obtained for different values of \( E_b/N_0 \) in this scenario. The table compares the values obtained by simulating pairs of TB codes \( R = 1/2, 3/5 \) and \( 3/4 \) with the asymptotic results obtained from (36) through Theorem 2. We can see that there is a discrepancy between simulation and asymptotic analysis.
Figure 4. Enumerative + TB coding, $n = 100$, $k = 80$. Frame error rate for separate and two-class source-channel coding.

Table I

| $E_b/N_0$ (dB) | Simulation  | Asymptotic analysis |
|---------------|-------------|---------------------|
| 2 dB          | (0.5, 0.75) | (0.447, 0.475)     |
| 3 dB          | (0.6, 0.75) | (0.481, 0.522)     |
| 4 dB          | (0.6, 0.75) | (0.516, 0.569)     |

This effect can be due to the short block length considered or to the coarse granularity of the coding rates, since $R_i \in \{1/2, 3/5, 3/4\}$, $i = 1, 2$.

2) Moderate block length scenario ($k = 1000$, $n \approx 1000$):

Figure 5 shows the FER for an implementation using LDPC codes and iterative decoding. We use enumerative source coding and a family of quasi-cyclic (QC) LDPC codes as channel codes. In particular we consider a set of codes with 24-column base matrix and coding rates $R = 12/24, 13/24, \ldots, 16/24$. For constructing these parity-check matrices we used the optimization algorithm from [28]. The only exception is the code of rate $R = 18/24$ which is borrowed from [34, code A]. The decoding algorithm is stopped after 50 iterations of belief propagation decoding. From Fig. 5 we can see that the proposed scheme outperforms separate coding by 0.4-0.7 dB. In this case the gap to the lower bound (46) is larger with respect to that in Fig. 4 because of the suboptimal decoding algorithm, with performance far from ML decoding.

Table II shows the best code rate pairs in this scenario. Given the granularity of the codes used, the agreement between asymptotic results and simulation results is surprisingly accurate. This is due to the larger block length, that makes the asymptotic approximations more accurate. This fact justifies the use of the asymptotic analysis from Section III to guide the design of good finite length codes.
Figure 5. Enumerative + LDPC coding, $n = 1008$, $k = 1000$. Frame error rate for separate and two-class source-channel coding.

Table II

| $E_b/N_0$ (dB) | Simulation | Asymptotic analysis |
|---------------|------------|---------------------|
| 1 dB          | (0.5, 0.6) | (0.499, 0.511)     |
| 2 dB          | (0.5, 0.6) | (0.536, 0.561)     |
| 3 dB          | (0.542, 0.583) | (0.575, 0.612)   |
| 4 dB          | (0.583, 0.667) | (0.614, 0.664)   |

V. Concluding Remarks

In this paper we have presented a source-channel coding scheme in which the source messages are divided into classes and a channel code and ML decoding is used for each of the classes. We have shown that the overall scheme outperforms separate source-channel coding and approaches the performance of joint source-channel coding as the number of classes increases.

The proposed scheme can be implemented using existing source and channel codes with reduced complexity. While the theoretical analysis assumed here memoryless sources and channels, the scheme can be implemented for sources and channels with memory by using appropriate source and channel codes. Simulation results for a binary memoryless source transmitted over a binary input additive Gaussian channel show that the proposed scheme with two classes offers a 0.5-1.0 dB gain compared to separate source-channel coding. This is consistent with the theoretically predicted values. Moreover, analytical results have been shown to offer a practical guideline to the design of finite-length source-channel codes in the memoryless setting.
APPENDIX I

PROOF OF LEMMA I

For each \( \rho \geq 0 \) and \( k = 1, 2, \ldots \), let us define the random variable \( Z_{\rho,k} \equiv \log P^k(V) \) with underlying distribution

\[
P^k_v \equiv \frac{P^k_v(v)^{1+\rho}}{\sum \bar{v} P^k(\bar{v})^{1+\rho}}.
\]

This distribution is the multi-letter version of (15). The asymptotic normalized log-moment generating function of \( Z_{\rho,k} \) is given by

\[
\kappa_{\rho}(\tau) \equiv \lim_{k \to \infty} \frac{1}{k} \log \mathbb{E}[e^{\tau Z_{\rho,k}}] = \log \left( \sum_v P(v)^{1+\rho+\tau} \right) - \frac{1}{1+\rho} E_s(\rho).
\]

From the definitions (12) and (13) we have that

\[
E_{s,i}(\rho) = \lim_{k \to \infty} \frac{1}{k} \log \left( \sum_{v \in A_i^k} P^k(v)^{1+\rho} \right) = \lim_{k \to \infty} \frac{1}{k} \log \left( \sum_{v \in \bar{A}_i^k} P^k(v)^{1+\rho} \right) = E_s(\rho) + (1+\rho) \lim_{k \to \infty} \frac{1}{k} \log \left( \Pr\left\{ \log \gamma^k < Z_{\rho,k} \leq \log \gamma^k_{i+1} \right\} \right).
\]

Applying the Gartner-Ellis theorem [35, Th. II.6.1] to the term \( \Pr\left\{ \log \gamma^k_i < Z_{\rho,k} \leq \log \gamma^k_{i+1} \right\} \) we obtain

\[
\Lambda_i(\rho) = E_{s,i}(\rho) \leq \overline{\Lambda}_i(\rho),
\]

where the rate functions \( \Lambda_i(\rho) \) and \( \overline{\Lambda}_i(\rho) \) are respectively given by

\[
\Lambda_i(\rho) \equiv \sup_{\log \gamma_i < r < \log \gamma_{i+1}} \inf_{\tau} \left\{ E_s(\rho) - (1+\rho)(r\tau - \kappa_{\rho}(\tau)) \right\},
\]

\[
\overline{\Lambda}_i(\rho) \equiv \sup_{\log \gamma_i \leq r \leq \log \gamma_{i+1}} \inf_{\tau} \left\{ E_s(\rho) - (1+\rho)(r\tau - \kappa_{\rho}(\tau)) \right\}.
\]

Given the smoothness properties of \( \kappa_{\rho}(\tau) \) both lower and upper bounds coincide and

\[
E_{s,i}(\rho) = \Lambda_i(\rho) = \overline{\Lambda}_i(\rho) = \Lambda_i(\rho).
\]

We compute now \( \Lambda_i(\rho) \). Using (51), the objective of either (56) or (57) can be written as

\[
\Phi(r,\tau) \equiv -(1+\rho)r\tau + \log \left( \sum_v P(v)^{1+\rho+\tau} \right)^{1+\rho}.
\]

It can be easily checked that the function \( \Phi(r,\tau) \) is differentiable in \( C^2 \) and that its Hessian is given by

\[
\nabla^2 \Phi(r,\tau) = \begin{bmatrix}
0 & -(1+\rho) \\
-(1+\rho) & \frac{\partial^2 \Phi(r,\tau)}{\partial \tau^2}
\end{bmatrix}.
\]
Hence, its determinant is $|\nabla^2 \Phi(r, \tau)| = -(1 + \rho)^2 < 0$ and the solution of (56)-(57) is a saddle point provided that the constraints are non-active. By taking the derivative of $\Phi(r, \tau)$ with respect to $\tau$ and equating it to zero we obtain that for the optimal point it holds that

$$r = \sum_v P_r(v) \log P(v),$$

(61)

where $\rho'$ is such that

$$\frac{1}{1 + \rho'} = \frac{1}{1 + \rho} + \tau.\quad(62)$$

Note that given (62) it is equivalent to perform the optimization over the domain of $\tau$ or over the domain of $\rho'$. By taking the derivative of $\Phi(r, \tau)$ with respect to $r$ and equating it to zero it follows that for the optimal point

$$\tau = 0 \Leftrightarrow \rho = \rho',$$

(63)

provided that the constraints are non-active.

We translate now the constraints on $r$ to the $\rho'$ domain. Using (61) and the definition of $\rho^*_i$ in Lemma 1, the constraints $\log \gamma_i \leq r \leq \log \gamma_{i+1}$ (resp. $\log \gamma_i < r < \log \gamma_{i+1}$) can be equivalently written as $\rho^*_{i+1} \leq \rho' \leq \rho^*_i$ (resp. $\rho^*_{i+1} < \rho' < \rho^*_i$), $i = 0, \ldots, N$.

We focus now on the solution of (57). Since the unconstrained solution to the optimization with respect to $\rho'$ is $\rho' = \rho$, we consider three regions based on the value of $\rho$:

1) When $\rho^*_{i+1} \leq \rho \leq \rho^*_i$ the constraints are non-active and the saddlepoint occurs at

$$r = \sum_v P_r(v) \log P(v), \quad \tau = 0.$$

(64)

Substituting these values into (59) we obtain

$$\Lambda_i(\rho) = E_i(\rho).$$

(65)

2) For $\rho > \rho^*_i$, the optimal $r$ is given by

$$r = \log \gamma_i = \sum_v P_r(v) \log P(v).$$

(66)

From (62) we obtain that

$$\tau = \frac{1}{1 + \rho^*_i} - \frac{1}{1 + \rho}$$

(67)

$$= \frac{\rho - \rho^*_i}{(1 + \rho^*_i)(1 + \rho)}.\quad(68)$$

Substituting these values into (59), it follows that

$$\Lambda_i(\rho) = \left\{ \begin{array}{c} \rho - \rho^*_i \log \gamma_i + \log \left( \sum_v P(v)^{1+\rho} \right)^{1+\rho} \\
 \end{array} \right\}$$

(69)

$$= \left\{ \begin{array}{c} \rho - \rho^*_i \log \gamma_i + \log \left( \sum_v P(v)^{1+\rho^*_i} \right)^{1+\rho^*_i} \\
 + \rho - \rho^*_i \log \left( \sum_v P(v)^{1+\rho^*_i} \right)^{1+\rho^*_i} \\
 \end{array} \right\}$$

(70)

$$= \left\{ \begin{array}{c} \rho - \rho^*_i \log \gamma_i + E_i(\rho^*_i) \\
 + E_i(\rho^*_i) \\
 \end{array} \right\},$$

(71)
where in (70) we added and subtracted the term \( \rho_i^* \log \left( \sum_v P(v)^{1+\rho_i^*} \right) \); and (71) follows from the definition of \( E_i(\rho) \).

3) Proceeding in an analogous way to the previous case, for \( \rho < \rho_{i+1}^* \), we obtain that

\[
\Lambda_i(\rho) = \left\{ (\rho - \rho_{i+1}^*) \frac{E_i(\rho_{i+1}^*) - \log \gamma_{i+1}}{1 + \rho_{i+1}^*} + E_i(\rho_i^*) \right\}.
\]

(72)

If \( \rho_{i+1}^* = 0 \) or \( \rho_i^* = \infty \), the first and last regions of the RHS of (16) are empty, respectively. Otherwise, using

\[
E'_i(\rho) = E_i(\rho) - \sum_v P_v(\rho) \log P(v)
\]

and (17), it follows that

\[
E'_i(\rho_i^*) = E_i(\rho_i^*) - \log \gamma_i.
\]

(74)

The result follows from substituting (65), (69)-(71) and (72), \( i = 0, \ldots, N \), into the corresponding range of the parameter \( \rho \), and using (74).

**APPENDIX II**

**Slope of \( E_{s,i}(\rho) \) as \( \rho \to \infty \)**

Using the characterization in Lemma 1 it follows that

\[
\lim_{\rho \to \infty} \frac{1}{\rho} E_{s,i}(\rho) = \lim_{\rho \to \infty} \frac{1}{\rho} \left( E_i(\rho_i^*) + (\rho - \rho_i^*) E'_i(\rho_i^*) \right)
\]

\[
= E'_i(\rho_i^*),
\]

(75)

(76)

as long as \( \rho_i^* < \infty \).

Also, using the definitions (12) and (13) we have that

\[
\lim_{\rho \to \infty} \frac{1}{\rho} E_{s,i}(\rho) = \lim_{\rho \to \infty} \lim_{k \to \infty} \frac{1}{\rho k} \log \left( \sum_{v \in A_k^i} P^k(v)^{1+\rho} \right)
\]

\[
= \lim_{k \to \infty} \lim_{\rho \to \infty} \frac{1}{\rho k} \log \left( \sum_{v \in A_k^i} P^k(v)^{1+\rho} \right)
\]

\[
= \lim_{k \to \infty} \frac{1}{k} \log |A_k^i|
\]

\[
= \frac{R_i}{t^i},
\]

(77)

(78)

(79)

(80)

(81)

where in (78) we applied the Moore-Osgood theorem [36, p. 619] since the expression

\[
\frac{1}{\rho k} \log \left( \sum_{v \in A_k^i} P^k(v)^{1+\rho} \right)^{1+\rho}
\]

(82)
presents uniform convergence for each \( k \) as \( \rho \to \infty \), and pointwise convergence as \( k \to \infty \), as we show next.

Then, using (75)-(76) and (77)-(81), it follows that the expressions (18) and (19) are equivalent.

In order to show the convergence properties of (82), note that

\[
\frac{1}{k} \log \left( \sum_{v \in A_k^i} P^k(v)^{1+\rho} \right)^{1+\rho} - \frac{1}{k} \log |A_k^i| \leq \frac{1}{k} \left( \log \left( \sum_{v \in A_k^i} 1^{1+\rho} \right)^{1+\rho} - \log |A_k^i| \right)
\]

\[
= \frac{1}{k} \left( \log |A_k^i|^{1+\rho} - \log |A_k^i| \right)
\]

\[
= \frac{1}{k\rho} \log |A_k^i|
\]

\[
= \frac{R_i}{1}.
\]

Similarly,

\[
\frac{1}{k} \log |A_k^i| - \frac{1}{k} \log \left( \sum_{v \in A_k^i} P^k(v)^{1+\rho} \right)^{1+\rho} \leq \frac{1}{k} \left( \log |A_k^i| - \log \left( \sum_{v \in A_k^i} \min P(v)^{1+\rho} \right)^{1+\rho} \right)
\]

\[
= \frac{1}{k} \left( \log |A_k^i| - \log |A_k^i|^{1+\rho} - \log \left( \min \limits_{v} P(v) \right)^{1+\rho} \right)
\]

\[
= - \frac{1}{k\rho} \log |A_k^i| - \frac{1}{\rho} \log \min \limits_{v} P(v)
\]

\[
= \frac{1}{\rho} \left( - \log \min \limits_{v} P(v) - \frac{R_i}{1} \right).
\]

Since (86) and (90) do not depend on \( k \), (82) presents uniform convergence with respect to \( k \) as \( \rho \to \infty \). Pointwise convergence of (82) as \( k \to \infty \) follows from Lemma 1.

**APPENDIX III**

**PROOF OF THEOREM 1**

Under our assumption that the number of classes \( N_k \) behaves sub-exponentially in \( k \), the error exponent is given by the minimum of the individual exponents of each of the summands in (33), namely

\[
- \lim_{n \to \infty} \frac{1}{n} \log \epsilon_n = \min \left\{ - \lim_{n \to \infty} \frac{1}{n} \log \Pr\{ V \in A_0^i \}, \min_{i=1, \ldots, N} - \lim_{n \to \infty} \frac{1}{n} \log \Pr\{ V \in A_k^i, \hat{V}_i \neq V \}, \right.

\[
- \lim_{n \to \infty} \frac{1}{n} \log \Pr\{ V \notin A_0^k, q(V, Y) \leq \max_{v \neq V, v \in A_0^k} q(v, Y) \} \right\}.
\]

We next analyze each of the terms in the minimum separately.

An application of (20)-(21) directly gives the exponent of the first term in the minimum, that is

\[
- \lim_{n \to \infty} \frac{1}{n} \log \Pr\{ V \in A_0^k \} = \rho_1^* R_1 - t E_s(\rho_1^*)
\]

\[
= \frac{R_1}{t}.
\]

We now upper bound the second term in (91). First, we use the chain rule to express the probability as

\[
\Pr\{ V \in A_k^i, \hat{V}_i \neq V \} = \Pr\{ \hat{V}_i \neq V | V \in A_k^i \} \Pr\{ V \in A_k^i \}, \ i = 1, \ldots, N,
\]
so as to estimate the exponent of each of the two factors separately. The first factor corresponds to the error probability of a channel coding problem with $M_i$ messages transmitted over a channel $W$. We can lower bound its exponent in terms of the random-coding exponent for input distribution $Q_i$. For each class $A_i^k$, $i = 1, \ldots, N$, there exists a code $C_i$ whose error probability over the memoryless channel $W$ satisfies \cite[Th. 5.6.1]{2}

$$-\lim_{n \to \infty} \frac{1}{n} \log \Pr\{\hat{V}_i \neq V | V \in A_i^k\} \geq \max_{\rho_i \in [0,1]} \left\{ E_0(\rho_i, W, Q_i) - \rho_i R_i \right\},$$

(95)

$$= E_r(R_i, Q_i).$$

(96)

Similarly to (93), the exponent of the second factor in (94) is given by

$$-\lim_{n \to \infty} \frac{1}{n} \log \Pr\{V \in A_i^k\} = t e\left(\frac{R_{i+1}}{R_i}\right).$$

(97)

Combining (96) and (97) we thus obtain

$$-\lim_{n \to \infty} \frac{1}{n} \log \Pr\{V \in A_i^k, \hat{V}_i \neq V\} \geq E_r(R_i, Q_i) + t e\left(\frac{R_{i+1}}{R_i}\right), \quad i = 1, \ldots, N.$$  

(98)

Finally, we identify the last term in (33) as the error exponent of a specific joint source-channel coding problem, where the source message probabilities do not add up to 1. In the random-coding argument, codewords are generated according to a class-dependent input distribution $Q_i$, $i = 1, \ldots, N$. We can thus use \cite[Th. 1]{17} to bound the exponent

$$-\lim_{n \to \infty} \frac{1}{n} \log \Pr\left\{q(V, Y) \leq \max_{\hat{V} \neq V, \hat{V} \notin A_0^i} q(\hat{v}, Y), V \notin A_i^k \right\} \geq \min_{i=1,\ldots,N} \left\{ E_0(\rho_i', W, Q_i) - t E_{s,i}(\rho_i') \right\},$$

(99)

for any $\rho_i' \in [0,1]$. Here we used that the proof of \cite[Th. 1]{17} is valid also for defective source message probabilities.

From Lemma 1, we infer that the source function $E_{s,i}(\rho)$ is non-decreasing, convex and with a non-decreasing derivative. Moreover, (19) shows that the derivative approaches the limiting value $\frac{R_i}{t}$ as $\rho \to \infty$. Therefore, the source function $E_{s,i}(\rho)$ satisfies the following simple upper bound for non-negative $\rho$

$$E_{s,i}(\rho) \leq E_{s,i}(0) + \rho \frac{R_i}{t},$$

(100)

$$= -e\left(\frac{R_{i+1}}{t}\right) + \rho \frac{R_i}{t},$$

(101)

where we used (21). Substituting (101) in the right-hand side (99) we obtain

$$E_0(\rho_i', W, Q_i) - t E_{s,i}(\rho_i') \geq E_0(\rho_i', W, Q_i) - \rho_i R_i + t e\left(\frac{R_{i+1}}{t}\right).$$

(102)

Since this inequality holds for arbitrary $\rho_i' \in [0,1]$ and input distribution $Q_i$, we conclude that for each value of $i = 1, \ldots, N$, the corresponding exponent in (99) is upper bounded by the exponent in (98).

Finally, we observe that the null-symbol input distribution $Q_0$ satisfies $E_r(R, Q_0) = 0$ for any rate $R$. Then, using the intermediate results (93), with $t e\left(\frac{R_i}{t}\right)$ replaced by $t e\left(\frac{R_i}{t}\right) + E_r(R_1, Q_0)$, and (98) we get the desired

$$-\lim_{n \to \infty} \frac{1}{n} \log \epsilon_n \geq \min_{i=0,\ldots,N} \left\{ E_r(R_i, Q_i) + t e\left(\frac{R_{i+1}}{t}\right) \right\}.$$  

(103)
APPENDIX IV
PROOF OF THEOREM 2

We start by writing (34) in dual form, that is, as explicit maximizations over parameters $\rho_i$ and $\rho_i'$,

$$
- \lim_{n \to \infty} \frac{1}{n} \log \epsilon_n \geq \min_{i=0,\ldots,N} \left\{ \max_{\rho_i \in [0,1]} \left\{ E_0(\rho_i, Q_i) - \rho_i R_i \right\} + \max_{\rho_i' \in [0,\infty]} \left\{ \rho_i' R_{i+1} - t E_s(\rho_i') \right\} \right\}.
$$

(104)

For $i = 0$ we have $E_s(R, Q_0) = 0$ and for $i = N$ we have $t E_s(R_N, \rho_i') = 0$. In the range $i = 1, \ldots, N - 1$ we may fix $\rho_i' = \rho_i$ without violating the inequality in (104). Finally, optimizing over $Q_i$, $i = 1, \ldots, N$, from (104) we obtain

$$
- \lim_{n \to \infty} \frac{1}{n} \log \epsilon_n \geq \max_{R_1 \geq \ldots \geq R_N \geq 0} \min_{\rho_0' \in [0,\infty]} \left\{ \max_{i=1,\ldots,N-1} \max_{\rho_i \in [0,1]} \left\{ E_0(\rho_i) - t E_s(\rho_i) - \rho_i (R_i - R_{i+1}) \right\} \right\},
$$

$$
\max_{\rho N \in [0,1]} \left\{ E_0(\rho_N) - \rho N R_N \right\}.
$$

(105)

Noting that the inner minimization in (105) is maximized with respect to $\{R_i\}$ when $R_i - R_{i+1}$ is constant, $i = 1, \ldots, N - 1$, the result follows.

REFERENCES

[1] C. Shannon, “A mathematical theory of communication,” Bell Syst. Tech. J., vol. 27, pp. 379–423 and 623–656, July and Oct. 1948.
[2] R. G. Gallager, Information Theory and Reliable Communication. New York: John Wiley & Sons, Inc., 1968.
[3] I. Csiszár, “Joint source-channel error exponent,” Probl. Contr. Inf. Theory, vol. 9, pp. 315–328, 1980.
[4] D. Wang, A. Ingber, and Y. Kochman, “The dispersion of joint source-channel coding,” in 49th Annual Allerton Conf. on Comm., Control, and Computing, Sept 2011, pp. 180–187.
[5] V. Kostina and S. Verdú, “Lossy joint source-channel coding in the finite blocklength regime,” IEEE Trans. Inf. Theory, vol. 59, no. 5, pp. 2545–2575, May 2013.
[6] Y. Zhong, F. Alajaji, and L. L. Campbell, “On the joint source-channel coding error exponent for discrete memoryless systems,” IEEE Trans. Inf. Theory, vol. 52, no. 4, pp. 1450–1468, Apr. 2006.
[7] J. Hagenauer, “Source-controlled channel decoding,” IEEE Trans. Commun., vol. 43, no. 9, pp. 2449–2457, 1995.
[8] J. Garcia-Frias and Y. Zhao, “Compression of binary memoryless sources using punctured turbo codes,” IEEE Comm. Lett., vol. 6, no. 9, pp. 394–396, 2002.
[9] M. Fresia, F. Pérez-Cruz, H. Poor, and S. Verdú, “Joint source and channel coding,” IEEE Signal Process. Mag., vol. 27, no. 6, pp. 104–113, Nov. 2010.
[10] G. Shamir, J. Boutros, A. Alloum, and L. Wang, “Non-systematic ldpc codes for redundant data,” in 2006 Inf. Theory and App. Inaugural Workshop, San Diego, CA, US, Feb. 2006.
[11] R. Bauer and J. Hagenauer, “Symbol-by-symbol MAP decoding of variable length codes,” ITG Fachbericht, pp. 111–116, 2000.
[12] J. Hagenauer and R. Bauer, “The turbo principle in joint source channel decoding of variable length codes,” in 2001 IEEE Inf. Theory Workshop, 2001, pp. 33–35.
[13] A. Guyader, E. Fabre, C. Guillemot, and M. Robert, “Joint source-channel turbo decoding of entropy-coded sources,” IEEE Journal on Sel. Areas in Comm., vol. 19, no. 9, pp. 1680–1696, Sep 2001.
[14] M. Grantet, P. Cosman, and G. Olmo, “Joint source/channel coding and map decoding of arithmetic codes,” IEEE Trans. on Comm., vol. 53, no. 5, pp. 905–905, May 2005.
[15] S. Lonardi, W. Szpankowski, and M. Ward, “Error resilient LZ’77 data compression: Algorithms, analysis, and experiments,” IEEE Trans. Inf. Theory, vol. 53, no. 5, pp. 1799–1813, May 2007.
[16] Y. Y. Shkel, V. Y. F. Tan, and S. C. Draper, “On mismatched unequal message protection for finite block length joint source-channel coding,” in *2014 IEEE Int. Symp. on Inf. Theory*, Honolulu, USA, June-July 2014.

[17] A. Tauste Campo, G. Vazquez-Vilar, A. Guillén i Fàbregas, T. Koch, and A. Martinez, “A derivation of the source-channel error exponent using non-identical product distributions,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 3209–3217, June 2014.

[18] F. Jelinek, *Probabilistic Information Theory*. New York: McGraw-Hill, 1968.

[19] A. Tauste Campo, G. Vazquez-Vilar, A. Guillén i Fàbregas, T. Koch, and A. Martinez, “Achieving Csiszár’s exponent for joint source-channel coding with product distributions,” in *2012 IEEE Int. Symp. on Inf. Theory*, Boston, USA, July 2012.

[20] I. Csiszár, “The method of types,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

[21] I. E. Bocharova, A. Guillén i Fàbregas, B. D. Kudryashov, A. Martinez, A. Tauste Campo, and G. Vazquez-Vilar, “Source-channel coding with multiple classes,” in *2014 IEEE Int. Symp. on Inf. Theory*, Honolulu, USA, June-July 2014.

[22] R. T. Rockafellar, *Convex Analysis*. Princeton, US: Princeton University Press, 1970.

[23] T. Cover, “Enumerative source encoding,” *IEEE Trans. Inf. Theory*, vol. 19, no. 1, pp. 73–77, Jan. 1973.

[24] I. H. Witten, R. M. Neal, and J. G. Cleary, “Arithmetic coding for data compression,” *Communications of the ACM*, vol. 30, no. 6, pp. 520–540, 1987.

[25] I. E. Bocharova, A. Guillén i Fàbregas, B. D. Kudryashov, A. Martinez, A. Tauste Campo, and G. Vazquez-Vilar, “Low-complexity fixed-to-fixed joint source-channel coding,” in *8th Int. Symp. on Turbo Codes and Iterative Inform. Process.*, Bremen, Germany, Aug. 2014.

[26] F. Hug, I. E. Bocharova, R. Johannesson, and B. D. Kudryashov, “Searching for high-rate convolutional codes via binary syndrome trellises,” in *2009 IEEE Int. Symp. Inf. Theory*, Seoul, Korea, Jun.–Jul. 2009.

[27] I. E. Bocharova and B. D. Kudryashov, “Rational rate punctured convolutional codes for soft-decision Viterbi decoding,” *IEEE Trans. Inf. Theory*, vol. 43, no. 4, pp. 1305–1313, Jul. 1997.

[28] I. Bocharova, B. Kudryashov, and R. Johannesson, “Combinatorial optimization for improving QC LDPC codes performance,” in *2013 IEEE Int. Symp. Inf. Th.*, Istanbul, Turkey, Jul. 2013.

[29] C. E. Shannon, “Probability of error for optimal codes in a Gaussian channel,” *Bell Syst. Tech. J.*, vol. 38, no. 3, pp. 611–656, May 1959.

[30] I. Sason and S. Shamai (Shitz), *Performance analysis of linear codes under maximum-likelihood decoding: a tutorial*. Foundations and Trends Commun. and Inf. Theory, now Publishers, 2006.

[31] S. Vialle and J. Boutros, “Performance of optimal codes on Gaussian and Rayleigh fading channels: a geometrical approach,” in *37th Ann. Allerton Conf. Comm., Contr. and Comp.*, Allerton, IL, USA, Sep. 1999.

[32] I. E. Bocharova, R. Johannesson, B. D. Kudryashov, and P. Stuhl, “Tailbiting codes: Bounds and search results,” *IEEE Trans. Inf. Theory*, vol. 48, no. 1, pp. 137–148, Jan. 2002.

[33] I. E. Bocharova, R. Johannesson, B. D. Kudryashov, and M. Lončar, “BEAST decoding for block codes,” *Eur. Tran. Tel.*, vol. 15, no. 4, pp. 297–305, Jul.–Aug. 2004.

[34] *Air Interface for Fixed and Mobile Broadband Wireless Access Systems*, IEEE P802.16e/D12 Draft, Oct. 2005.

[35] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*. Berlin, Germany: Springer, 1985, vol. 271.

[36] W. F. Osgood, *Lehrbuch der Funktionentheorie*, 5th ed. Leipzig: Teubner, 1928, vol. I.