Problems of Information Transmission, Vol. 37, No. 1, 2001, pp. 10–23. Translated from Problemy Peredachi Informatsii, No. 1, 2001, pp. 12–27.

Joint Source–Channel Coding With Variable Length Codes

V. B. Balakirsky

Received August 11, 1999; in final form, June 13, 2000

Abstract—We address the problem of joint source–channel coding when variable length codes are used for information transmission over a discrete memoryless channel. The data transmitted over the channel are interpreted as pairs (m_k, t_k), where m_k is the message generated by the source and t_k is the time instant when the transmission of the k–th codeword begins. The decoder constructs an estimate of the transmitted sequence of pairs, and the k–th decoding error is introduced as the event that the pair (m_k, t_k) does not belong to this sequence. We describe the maximum likelihood decoding algorithm and prove a lower bound on the exponent of decoding error probability. For a subclass of discrete memoryless sources and discrete memoryless channels this bound is asymptotically tight.

1. INTRODUCTION AND STATEMENT OF THE PROBLEM

Suppose that there is a discrete memoryless source generating a sequence of messages $\boldsymbol{m} = (m_1, m_2, \ldots)$, where $m_i \in [M] = \{1, \ldots, M\}$ for all $i \geq 1$. For all $k \geq 1$, let $\boldsymbol{m}_k = (m_1, \ldots, m_k)$. Then the probability that the source generates the messages included in the vector $\boldsymbol{m}_k$ can be expressed as

$$P(\boldsymbol{m}_k) = \prod_{i=1}^{k} P_{m_i},$$

where $P = \{P_m, m \in [M]\}$ is a given probability distribution. If $M = |\mathcal{U}^n|$, where $\mathcal{U}$ is a finite set, then we write $m_i = (m_i^{(1)}, \ldots, m_i^{(n)}) \in \mathcal{U}^n$, $i = 1, 2, \ldots$

Suppose that the sequence $\boldsymbol{m}$ enters the encoder, who constructs the code sequence $\boldsymbol{x}(\boldsymbol{m}) = (\boldsymbol{x}(m_1), \boldsymbol{x}(m_2), \ldots)$ concatenating codewords of a given variable length code. The code is defined by the mapping

$$m \in [M] \longrightarrow \boldsymbol{x}(m) \in \mathcal{X}^{\ell(m)},$$

where $\mathcal{X}$ is a finite set. We will specify $\ell(1), \ldots, \ell(M)$ by the vector $\ell = (\ell(1), \ldots, \ell(M))$ and call this vector the length distribution of the code. We also denote

$$\mathcal{L} = \bigcup_{m \in [M]} \Big\{ \ell(m) \Big\} \tag{1}$$

and

$$l_{\min} = \min_{m \in [M]} \ell(m), \quad l_{\max} = \max_{m \in [M]} \ell(m). \tag{2}$$

Furthermore, for all $l \in \mathcal{L}$, let

$$a_l = \sum_{m \in [M]} \chi\Big\{ \ell(m) = l \Big\} \tag{3}$$

denote the number of codewords of length l; hereafter χ stands for the indicator function : $\chi\{\mathcal{S}\} = 1$ if the statement $\mathcal{S}$ is true and $\chi\{\mathcal{S}\} = 0$ otherwise. Then

$$\bar{l} = \sum_{m \in [M]} P_m \ell(m) \tag{4}$$

is the average codeword length.

Suppose that we are given a discrete memoryless channel defined by the conditional probabilities $W(y|x), x \in \mathcal{X}, y \in \mathcal{Y}$, where $\mathcal{Y}$ is a finite set. Then, for all $t \geq 1$ and any sequences $\boldsymbol{x}_t = (x_1, \dots, x_t)$ and $\boldsymbol{y}_t = (y_1, \dots, y_t)$, the probability to receive $\boldsymbol{y}_t$ when $\boldsymbol{x}_t$ was sent is equal to

$$W(\boldsymbol{y}_t|\boldsymbol{x}_t) = \prod_{j=1}^{t} W(y_j|x_j).$$

The sequence $\boldsymbol{x}(\boldsymbol{m})$ is transmitted over the channel and the decoder constructs the decoded sequence of messages $\widehat{\boldsymbol{m}} = (\widehat{m}_1, \widehat{m}_2, \dots)$ having received the sequence $\boldsymbol{y}$. The pair of mappings (encoding, decoding) can be introduced as

$$\begin{array}{llcl} \widetilde{f}_{\rm sc}: & \boldsymbol{m} = (m_1, m_2, \dots) \in [M]^* & \longrightarrow & \boldsymbol{x}(\boldsymbol{m}) = (x_1, x_2, \dots) \in \mathcal{X}^*, \\ \widetilde{\varphi}_{\rm sc}: & \boldsymbol{y} = (y_1, y_2, \dots) \in \mathcal{Y}^* & \longrightarrow & \widehat{\boldsymbol{m}} = (\widehat{m}_1, \widehat{m}_2, \dots) \in [M]^* \end{array}$$

where the "sc" index emphasizes the point that we are dealing with a joint source–channel coding problem, $\sim$ shows that the codewords may have different lengths, and $*$ indicates that the length of the sequences is not specified. We will analyze the probability of the event called "the decoding error" and first present some motivations for this study.

1. Some difficulties with the definition the decoding error appear as soon as variable length codes are used. For example, suppose that these codes (say, Huffman codes) are used for source coding, the sequence at the output of the source encoder is split into blocks of a given length and encoded again using a block code for transmission over the noisy channel. Then the channel decoder can make an error leading to the loss of synchronization of the source decoder. As a result, "the catastrophic error propagation", can start. Note also that the probability of insertions and deletions of messages in the decoded sequence is positive for *any* decoding algorithm if the channel is noisy. In particular, we meet problem of classifying the case when $\widehat{m}_{i+1} = m_i$ for all $i \geq 2$ and $\widehat{m}_1, \widehat{m}_2 \neq m_1$. We will overcome these difficulties in the following way. Let us interprete data to be transmitted as the sequence of pairs $(m_1, t_1), (m_2, t_2), \dots$, where $t_1 = 0$ and $t_i = t_{i-1} + \ell(m_i)$ for all $i \geq 2$. In other words, the first component of each pair is the transmitted message and the second one is the time instant when the transmission of the corresponding codeword begins. The decoded sequence of messages can be represented similarly, and we say that the k–th decoding error takes place if and only if the decoded sequence of pairs $(\widehat{m}_1, \widehat{t}_1), (\widehat{m}_2, \widehat{t}_2), \dots$ does not contain the pair (m_k, t_k). We will analyze the average decoding error probability defined as the sum of the k–th decoding error probabilities on $k = 1, \dots, K$ divided by K, where K is fixed.

2. There is a class of problems when codes with the given length distribution have to be used by the statement of the problem, and the time instants when the transmission of the codewords begins are so important as the messages themselves. We will present a searching problem having these features.

3. There are known approaches to the decoding of variable length codes for data transmission over noisy channels that do not allow the decoding delay, i.e., the decoder simultaneously has to recognize the last symbol of the transmitted codeword and output the message, which is encoded by this codeword. Probably, these approaches can be hardly implemented in practice since the requirements to the characteristics of the code having a reasonable performance are very strong. We are interested in the possibility of reducing these requirements by introducing the decoding delay and attack the decoding problem using the methods developed for convolutional codes.

4. One of the directions of the study of the code performance relates to the exponent of the decoding error probability. If the source and channel coding problems are considered separately and the block codes of a fixed length are used, then the resulting exponent is equal to the minimum

between the exponents obtained at these two steps. However, there are results which show that one can achieve the sum of these two exponents instead of the minimum if the problems are considered jointly (this statement is presented precisely in the next section). We want to extend this approach to variable length codes and include the discussion of the final result below.

Suppose that the messages are binary vectors of length n and they are generated by the Bernoulli source defined by the probability p. Then the typical messages have the Hamming weight $w \approx np$. However, as it will follow from the results presented in Section 2, the lower bound on the exponent of the decoding error probability when block codes are used for joint source–channel coding is determined by the messages of the Hamming weight np_0, where $p_0 > p$. The variable lengths of the codewords are used to decrease the influence of these messages on the decoding error probability in such a way that the main contribution is caused by typical messages. In the proof of the theorem we show that if the length of a codeword for the message $m \in \mathcal{U}^n$ is assigned as

$$\ell(m) \approx \begin{cases} n, & \text{if } \operatorname{wt}(m) < np, \\ (-\log P_m)/h(p), & \text{if } \operatorname{wt}(m) \geq np, \end{cases} \tag{5}$$

where wt denotes the Hamming weight and $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function, then the average length of a codeword is close to n and the estimates of the decoding error probability have desired properties (hereafter, all logarithms are to the base 2).

The methods of analysis of joint source–channel coding problems from information theory point of view were developed in [1], [2], [3], and other papers. An overview of the results, which are directly relevant to foregoing considerations, is presented in Section 2. The first analysis of prefix codes for noisy channels (these codes form a subclass of variable length codes) was given in [4] where the authors considered the situation when the decoding delay is not allowed and the decoder is has to output the decoded message as soon as he receives the last symbol of corresponding codeword. The main difference of our considerations from this approach is that we will allow the decoding delay and consider decoding of the whole sequence of messages. The formal definition of the event, which we call "the decoding error", is given in Section 3 where we also extend the Viterbi–type algorithm [5] for convolutional codes to variable length codes. In Section 4 we illustrate the notion of the decoding error probability and describe the decoding algorithm for a class of search problems. In Section 5 we specially emphasize the point where the parameter, which plays the role of the rate of a variable length code, appears. The analysis of the exponent error probability in the ensemble of variable length codes, which is based on the random coding technique [5], [6], can be found in Section 6.

From mathematical point of view, the main result of the paper is the statement that $E_{\text{ran}}(H(P))$ is the asymptotic lower bound on the exponent of the decoding error probability for variable length codes, where $H(P)$ is the entropy of the source and $E_{\text{ran}}(\cdot)$ is the value of the exponent of the random coding bound for the channel. If the random coding bound for the channel at $H(P)$ is tight, then this lower bound is also tight. Constructing of the exponent is illustrated in Figure 1 for the Bernoulli source defined by the probability p.

2. SOME KNOWN RESULTS ON THE EXPONENT OF THE DECODING ERROR PROBABILITY FOR BLOCK CODES OF A FIXED LENGTH

A block code of length n for the pair (discrete memoryless source, discrete memoryless channel) is a pair of functions

$$f_{\text{sc}} : \mathcal{U}^n \to \mathcal{X}^n, \quad \varphi_{\text{sc}} : \mathcal{Y}^n \to \mathcal{U}^n,$$

known as encoding and decoding. The parameter

$$e_{\text{sc}} = \lim_{n\to\infty} \max_{(f_{\text{sc}},\varphi_{\text{sc}})} \frac{-\log \lambda_n(f_{\text{sc}},\varphi_{\text{sc}})}{n},$$

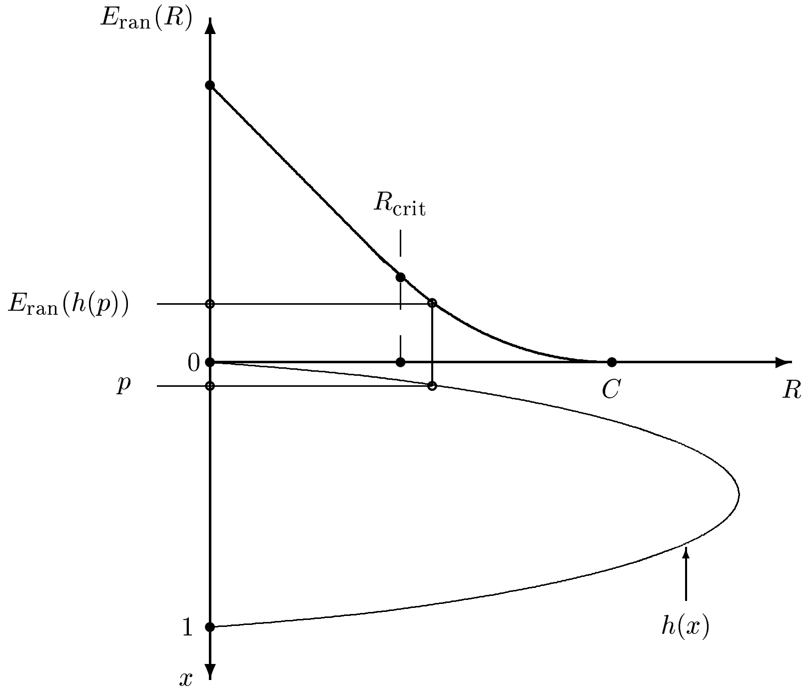

Fig. 1.

where

$$\lambda_n(f_{\rm sc},\varphi_{\rm sc}) = \sum_{m\in\mathcal{U}^n} P_m \sum_{\boldsymbol{y}_n\in\mathcal{Y}^n} W(\boldsymbol{y}_n|f_{\rm sc}(m))\chi\Big\{\ \varphi_{\rm sc}(\boldsymbol{y}_n)\neq m\ \Big\},$$

is referred to as the joint source–channel exponent of the decoding error probability. Note that the encoding $f_{\rm sc}$ and the decoding $\varphi_{\rm sc}$ are synchronized in a sense that as soon as the source outputs n symbols of the alphabet $\mathcal{U}$, the decoder receives n symbols of the alphabet $\mathcal{Y}$.

The lower bound given below is essentially better than the obvious bound

$$e_{\rm sc} \le \max_R \min\Big\{\ e_{\rm s}(R), e_{\rm c}(R)\ \Big\},$$

where $e_{\rm s}(R)$ and $e_{\rm c}(R)$ are the exponents for the source and channel coding at rate R.

Theorem 1 (The result was proved by Csiszár [2]; further extensions are presented in [3]). *Let C_0 denote the zero–error capacity of the channel.*

1. *If $C_0 > 0$, then*

$$e_{\rm sc} = \min_{R:\ R\ge H(P)} [\ e_{\rm s}(R) + e_{\rm c}(R)\ ]. \tag{6}$$

2. *If $C_0 = 0$, then*

$$E_{\rm sc} \le e_{\rm sc} \le \widehat{E}_{\rm sc}, \tag{7}$$

where

$$E_{\rm sc} = \min_{\widehat{P}:\ H(\widehat{P})\ge H(P)} [\ D(\widehat{P}\parallel P) + E_{\rm ran}(H(\widehat{P}))\ ],$$

$$\widehat{E}_{\rm sc} = \min_{\widehat{P}:\ H(\widehat{P})\ge H(P)} [\ D(\widehat{P}\parallel P) + E_{\rm sp}(H(\widehat{P}))\ ],$$

$\widehat{P} = \{\widehat{P}(u), u \in \mathcal{U}\}$ is a probability distribution on $\mathcal{U}$,

$$D(\widehat{P} \parallel P) = \sum_{u \in \mathcal{U}} \widehat{P}(u) \log \frac{\widehat{P}(u)}{P(u)}, \tag{8}$$

$$H(P) = -\sum_{u \in \mathcal{U}} P(u) \log P(u),$$

and

$$E_{\text{ran}}(R) = \max_{0 \le \rho \le 1} [\, E_0(\rho) - \rho R \,], \tag{9}$$

$$E_{\text{sp}}(R) = \max_{\rho \ge 0} [\, E_0(\rho) - \rho R \,] \tag{10}$$

are the values of the random coding bound and the sphere packing bound, respectively, expressed using the Gallager function

$$E_0(\rho) = -\log \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \omega(x) W^{1/(1+\rho)}(y|x) \right]^{1+\rho}$$

for some input probability distribution $\{\, \omega(x), x \in \mathcal{X} \,\}$.

Let us consider the Bernoulli source defined by the probability p and suppose that the messages are binary vectors of length n. We can construct $n+1$ codes $\mathcal{G}_0, \ldots, \mathcal{G}_n$ and if the message to be transmitted has weight w, encode it using the code $\mathcal{G}_w$. For the channels with $C_0 > 0$ we first transmit w with the decoding error probability zero and then decode the message in the code $\mathcal{G}_w$. A simple formal analysis of this algorithm leads to (6). However, a direct generalization of the algorithm to the channels with $C_0 = 0$ does not lead to the same result since the probability to decode the weight incorrectly is estimated as $2^{n' e_{\text{c}}(0)}$, where n' is the length of the code of cardinality $n+1$. If we want to obtain the same exponent of the total decoding error probability as before, we have to assign $n' = n e_{\text{sc}}(R^*)/e_{\text{c}}(0)$, where R^* is the parameter minimizing the expression on the right–hand side of (6). However $e_{\text{c}}(0) < \infty$, and this assignment linearly increases the total length of the codewords. Therefore the lower bound (7) was proved in [2] by simultaneous decoding of the weight and the message. Similar considerations for variable length codes lead to a conclusion that special short messages inserted into the stream of codewords to recover synchronization essentially decrease the exponent of the decoding error probability, and the synchronization problem cannot be separated from the decoding problem.

3. THE DECODING ERROR PROBABILITY FOR VARIABLE LENGTH CODES AND THE MAXIMUM LIKELIHOOD DECODING ALGORITHM

For a given length distribution of a variable length code ℓ, let us consider a non–uniform tree with M terminal nodes numbered $1, \ldots, M$ in such a way that the node m is located at level $\ell(m)$ (see Figure 2). The tree can be constructed in an arbitrary way in a sense that any number of edges can leave internal nodes. Let us assign the code symbols $x \in \mathcal{X}$ to every edge and assume that if the message m is generated by the source, then $\ell(m)$ symbols associated with the path leading to the node m enter the channel.

Suppose that the tree contains J internal nodes and associate them with integers $0, \ldots, J-1$ in such a way that 0 is assigned to the root. If we also associate 0 with all terminal nodes of the tree, then we obtain a code trellis (Figure 3) where at most J nodes are located at each level and there is a one–to–one mapping between sequences of messages and corresponding paths. Such a construction can be also viewed as the trellis of a punctured convolutional code [7], where the puncturing is understood in graph theory sense : we introduce a regular trellis and then remove some edges. Therefore the decoding algorithm described below can be represented as a variant of the Viterbi decoding for the punctured convolutional code. Note however that puncturing is rather peculiar in our case since competitions between different paths are possible only in the node 0 and since the number of competitors is at most M at each level.

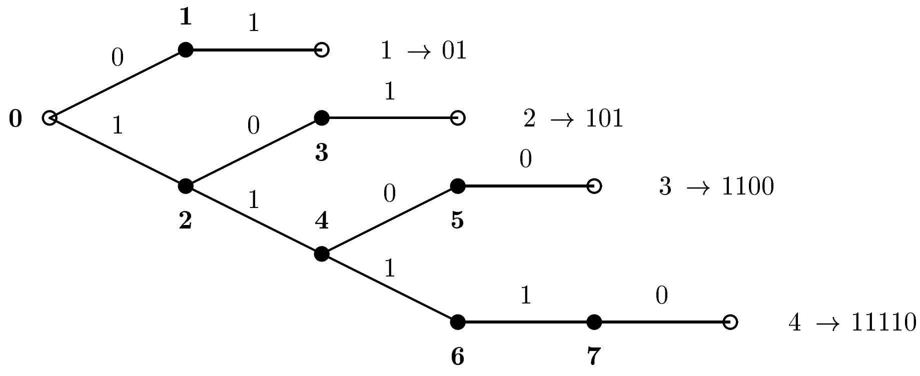

Fig. 2.

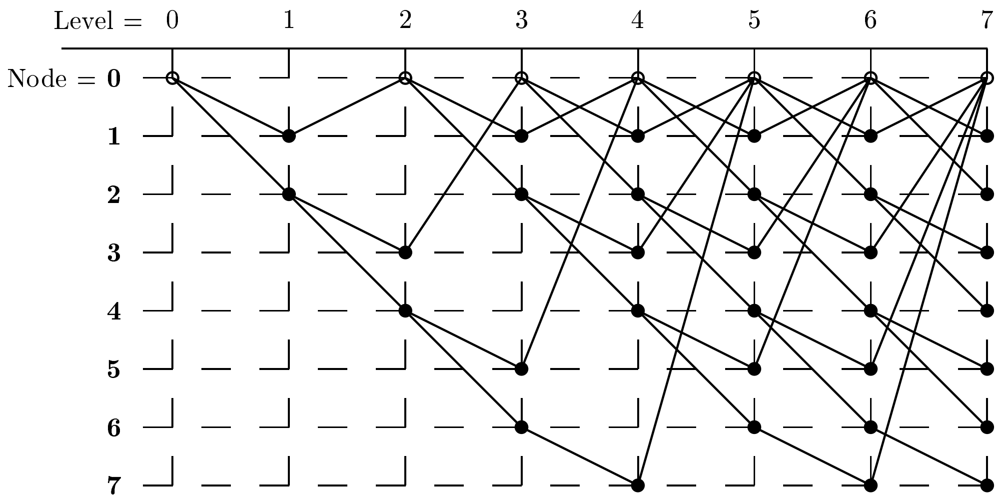

Fig. 3.

We will use the definition of the decoding error for variable length codes, which relates this event to code sequences.

Definition. *Given the sequence of messages* $\boldsymbol{m}$ *generated by the source and the decoded sequence of messages* $\widehat{\boldsymbol{m}}$, *let*

$$\mathcal{T}(\boldsymbol{m}) = \Big\{\, t_1, t_2, \dots \Big\}, \;\; \mathcal{T}(\widehat{\boldsymbol{m}}) = \Big\{\, \widehat{t}_1, \widehat{t}_2, \dots \Big\}, \tag{11}$$

where $t_1 = \widehat{t}_1 = 0$ *and, for all* $k, j = 2, 3, \dots,$

$$t_k = \ell(m_1) + \dots + \ell(m_{k-1}), \;\; \widehat{t}_j = \ell(\widehat{m}_1) + \dots + \ell(\widehat{m}_{j-1}) \tag{12}$$

are lengths of the sequences $(\boldsymbol{x}(m_1), \dots, \boldsymbol{x}(m_{k-1}))$ *and* $(\boldsymbol{x}(\widehat{m}_1), \dots, \boldsymbol{x}(\widehat{m}_{j-1}))$. *The* k*-th message* m_k *is decoded incorrectly, where* $k \in \{1, 2, \dots\}$, *if and only if one of the following conditions hold:*
E1. $t_k \notin \mathcal{T}(\widehat{\boldsymbol{m}})$;
E2. $t_k = \widehat{t}_j \in \mathcal{T}(\widehat{\boldsymbol{m}})$ *for some* $j \in \{1, 2, \dots\}$, *but* $\widehat{m}_j \neq m_k$.

Conditions E1 and E2 are illustrated in Figure 4. Note that if E1 is satisfied, then the decoder losses the time instant when the codeword for the k–th message begins. If E2 is satisfied, then the decoder recognizes this time instant, but incorrectly decodes the message.

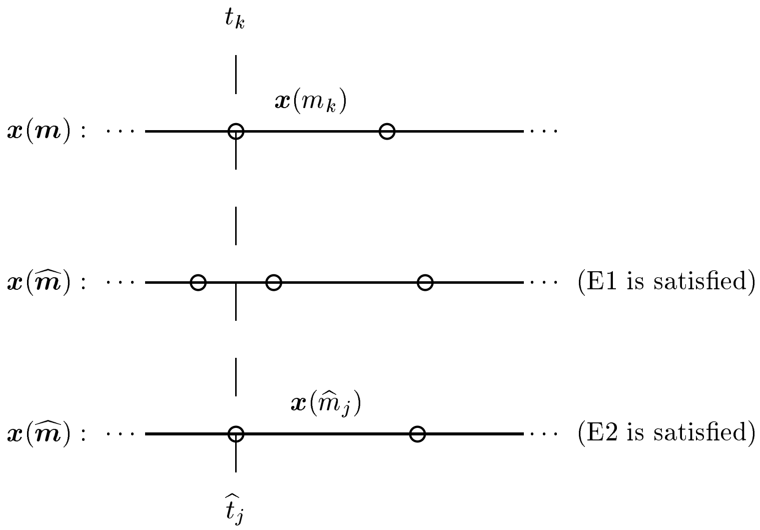

Fig. 4.

Let

$$\lambda_n^{(k)}(\widetilde{f}_{\rm sc}, \widetilde{\varphi}_{\rm sc} | \boldsymbol{m}, \ell) = \sum_{\boldsymbol{y}} W(\boldsymbol{y} | \boldsymbol{x}(\boldsymbol{m})) \sum_{\widehat{\boldsymbol{m}}} \chi \Big\{ \widehat{\boldsymbol{m}} = \widetilde{\varphi}_{\rm sc}(\boldsymbol{y}); \;\; \text{E1 or E2 is satisfied} \Big\} \tag{13}$$

for all $k = 1, 2, \dots$ Then

$$\Lambda_n^{(K)}(\widetilde{f}_{\rm sc}, \widetilde{\varphi}_{\rm sc} | \ell) = \frac{1}{K} \sum_{k=1}^{K} \sum_{\boldsymbol{m}} P(\boldsymbol{m}) \lambda_n^{(k)}(\widetilde{f}_{\rm sc}, \widetilde{\varphi}_{\rm sc} | \boldsymbol{m}, \ell) \tag{14}$$

is the average decoding error probability computed over the first K messages. Therefore the exponent of the decoding error probability can be introduced as

$$\widetilde{e}_{\rm sc}^{(K)} = \lim_{n \to \infty} \max_{\ell : \bar{l} \leq n} \max_{(\widetilde{f}_{\rm sc}, \widetilde{\varphi}_{\rm sc})} \frac{-\log \Lambda_n^{(K)}(\widetilde{f}_{\rm sc}, \widetilde{\varphi}_{\rm sc} | \ell)}{n}, \tag{15}$$

where we take the maximum over all length distributions of variable length codes for a given M such that the average codeword length is at most n and the maximum over all pairs (encoding, decoding) constructed for a fixed length distribution.

Suppose that $\boldsymbol{y}$ is the received sequence and, for all $t \geq 1$, let

$$\mathcal{X}_t = \Big\{ \boldsymbol{x}_t \in \mathcal{X}^t : \ \boldsymbol{x}_t = (\boldsymbol{x}(m_1), \ldots, \boldsymbol{x}(m_j)), \text{ for some } j \geq 1 \Big\} \tag{16}$$

denote the set consisting of code sequences of length t. If $\mathcal{X}_t \neq \varnothing$, then we set

$$\boldsymbol{x}_t^* = \arg \min_{\boldsymbol{x}_t \in \mathcal{X}_t} \Big[-\log W(\boldsymbol{y}_t | \boldsymbol{x}_t) \Big],$$

and

$$D(\boldsymbol{y}_t) = \begin{cases} -\log W(\boldsymbol{y}_t | \boldsymbol{x}_t^*), & \text{if } \mathcal{X}_t \neq \varnothing, \\ \infty, & \text{if } \mathcal{X}_t = \varnothing. \end{cases} \tag{17}$$

Then, as it is easy to see, $D(\boldsymbol{y}_1), D(\boldsymbol{y}_2), \ldots$ can be found from the recurrent equations

$$D(\boldsymbol{y}_t) = \min_{m \in [M]} \Big\{ D(\boldsymbol{y}_{t-\ell(m)}) - \sum_{l=1}^{\ell(m)} \log W(y_{l+t-\ell(m)} | x_l(m)) \Big\}, \quad t = 1, 2, \ldots, \tag{18}$$

where $D(\boldsymbol{y}_t) = \infty$ if $t < 0$ and $D(\boldsymbol{y}_0) = 0$. Equations (18) describe the maximum likelihood decoding in the following sense. Suppose that the decoder has the values $D(\boldsymbol{y}_{t-l_{max}}), \ldots, D(\boldsymbol{y}_{t-1})$ and corresponding sequences $\boldsymbol{x}_{t-l_{max}}^*, \ldots, \boldsymbol{x}_{t-1}^*$ in the memory at time instant t. Using these data he computes $D(\boldsymbol{y}_t)$ in accordance with (18) and constructs $\boldsymbol{x}_t^* = (\boldsymbol{x}_{t-\ell(\widehat{m})}^*, \boldsymbol{x}(\widehat{m}))$, where $\widehat{m}$ denotes the value of m minimizing the sum on the right–hand side of (18). As a result, the content of the memory is updated in such a way that now the decoder has $D(\boldsymbol{y}_{t-l_{max}+1}), \ldots, D(\boldsymbol{y}_t)$ and corresponding sequences $\boldsymbol{x}_{t-l_{max}+1}^*, \ldots, \boldsymbol{x}_t^*$.

Let us assume the following :

1) the total length of the code sequence transmitted over a channel is equal to T and known to the decoder;
2) the decoding delay is not important, and the decoder constructs an estimate of the whole sequence of messages having received the T–th symbol.

Then the decoding terminates at time instant T, and the decoder outputs the decoded sequence of messages, which is encoded by the code sequence $\boldsymbol{x}_T^*$.

Note that assumptions above are not essential and introduced to avoid the analysis of the so–called truncation decoding error probability [7]. In our case, the truncation would mean that the decoder constructs the estimate of the first message at time instant $\widetilde{t}$, the estimate of the second message at time instant $\widetilde{t} + \ell(\widehat{m}_1)$, etc. However, the value of the parameter $\widetilde{t}$ can be assigned as a linear function of n in such a way that the exponent of the decoding error probability caused by the truncation does not affect the asymptotic behaviour of the total decoding error probability. We skip further details because they are rather conventional for convolutional coding technique.

The maximum likelihood decoding maximizes the *a posteriori* probability of the sequence delivered to the receiver only if the probabilities of all sequences of messages that are mapped to code sequences belonging to the set $\mathcal{X}_t$ (see (16)), are equal to each other. To prove the random coding bound in Section 6, we assign the codewords lengths in such a way that this condition is satisfied and the values of t with $\mathcal{X}_t \neq \varnothing$ are big enough (this assignment for the Bernoulli source is shown in (5)). Nevertheless, the algorithm can be also used both in the case when the probability distribution on the messages is unknown and when the codewords length distribution is fixed. In particular, these situations can take place in the searching problems presented in the next section. Namely, we conclude that the maximum likelihood decoding is the optimum one for a fixed sequence of messages and fixed encoding function in this case.

4. HOW TO DECODE THE NOISY ANSWERS WHEN THE QUESTIONS ARE UNKNOWN

Let us consider the information transmission system in Figure 5 [8].

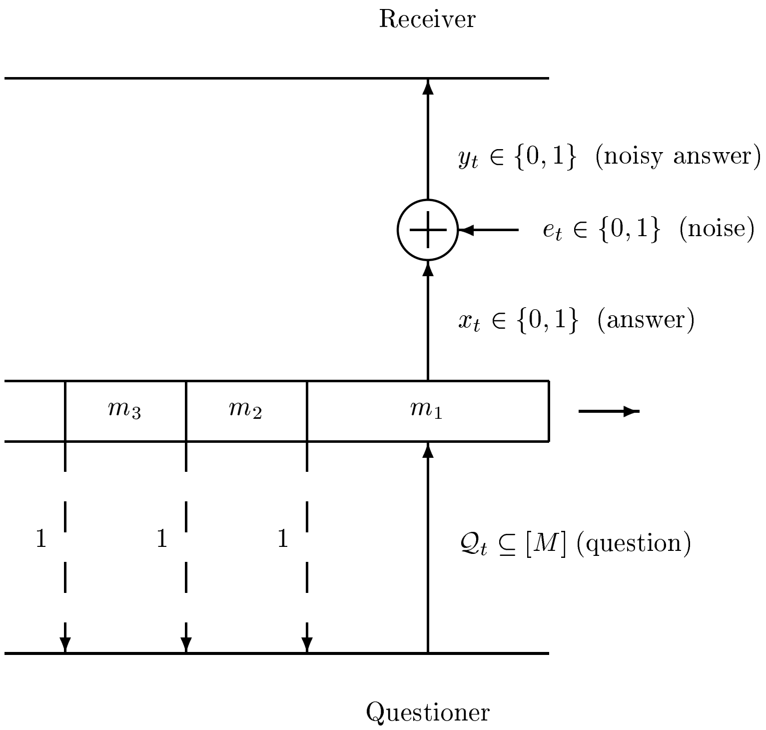

Fig. 5.

Suppose that there is a collection of boxes of M types numbered $1, \dots, M$. All boxes of type m have length $\ell(m)$ meters, $m = 1, \dots, M$, and the boxes of different types may have the same length. Somebody takes some box and puts it on the moving line, the speed of the line is 1 meter/second. Once per second, the questioner forms a question and sends it to the line. The question $\mathcal{Q}_t$ at the t–th second is a subset of the set $[M]$, and the answer $x_t = 1$ if the type of a box above the questioner belongs to $\mathcal{Q}_t$, and $x_t = 0$ otherwise. The answer goes to the receiver and can be corrupted with the noise, i.e., he gets $y_t = x_t \oplus e_t$, where $e_t \in \{0,1\}$. The goal of the receiver is to decode the sequence of types of the boxes that were put on the line.

Note that the questioner does not know the answer, while the receiver does not know the question. The only assumption, which we need to present a possible solution to the problem, is as follows : suppose that the questioner knows whether the current box is the previous or the new one. Formally, by the t–th second he receives the bit

$$b_t = \begin{cases} 0, & \text{if } t-1 \notin \mathcal{T}(\boldsymbol{m}), \\ 1, & \text{if } t-1 \in \mathcal{T}(\boldsymbol{m}), \end{cases}$$

where $\boldsymbol{m}$ is the sequence of types of the boxes and the set $\mathcal{T}(\boldsymbol{m})$ is defined in (11), and forms $\mathcal{Q}_t$ depending on b_t.

Suppose that participants know M and the set of lengths $\mathcal{L}$ in advance. Then they can construct a binary variable length code containing a_l codewords of length l for all $l \in \mathcal{L}$ (see (1)–(3)). After that they can construct a sequence $(\mathcal{Q}^{(1)}, \ldots, \mathcal{Q}^{(l_{max})})$, where

$$\mathcal{Q}^{(l)} = \Big\{ m \in [M] : \; \ell(m) \le l, \; x_l(m) = 1 \Big\}$$

and $x_l(m) \in \{0,1\}$ denotes the l–th component of the codeword corresponding to the m–th message, and the questioner can use the following algorithm : for all $t \ge 1$ (we assume that $b_1 = 1$) : if $b_t = 1$, then $i := 1$; if $b_t = 0$, then increase i by 1; $\mathcal{Q}_t := \mathcal{Q}^{(i)}$. Then, as it is easy to see, the sequence of answers is obtained as a concatenation of codewords associated with the types of the boxes on the line. The receiver can use the maximum likelihood decoding algorithm and construct the decoded sequence of types of the boxes. The event, called "the decoding error", makes sense for our search problem : if the E1 condition is satisfied, then the receiver losses the position where the new box has appeared; if the E2 condition is satisfied, then he knows this position, but losses the type of the box.

For example, suppose that $M = 4$, $\ell = (2,3,4,5)$, and that the questioner and receiver have constructed a code tree in Figure 2. In this case, the problem can be solved using the questions

$$\Big(\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)}, \mathcal{Q}^{(3)}, \mathcal{Q}^{(4)}, \mathcal{Q}^{(5)} \Big) = \Big(\{2,3,4\}, \{1,3,4\}, \{2,4\}, \{4\}, \varnothing \Big).$$

5. ESTIMATION OF THE NUMBER OF CODE SEQUENCES OF A GIVEN LENGTH

For all $t \ge 1$, let

$$N_t = |\, \mathcal{X}_t \,| \tag{19}$$

denote the number of code sequences of length t (see (16)). Then $N_t, t = 1, 2, \ldots$, can be found from the recurrent equations

$$N_t = \sum_{l \in \mathcal{L}} a_l N_{t-l}, \quad t = 1, 2, \ldots, \tag{20}$$

where $N_0 = 1$ and $N_t = 0$ for all $t \le 0$.

Proposition. *If r is a parameter satisfying the inequality*

$$\sum_{l \in \mathcal{L}} a_l 2^{-rl} \le 1, \tag{21}$$

then

$$N_t \le 2^{rt}, \quad \textit{for all} \;\; t \ge 1. \tag{22}$$

Proof. If (21) is valid then

$$2^{rt} \ge \sum_{l \in \mathcal{L}} a_l 2^{r(t-l)},$$

and subtracting (20) we obtain

$$\delta_t \ge \sum_{l \in \mathcal{L}} a_l \delta_{t-l},$$

where $\delta_t = 2^{rt} - N_t$. Since $\delta_t \ge 0$ for all $t \le 0$, these inequalities are also valid for all $t \ge 1$, as it can be easily checked using induction on t. $\triangle$

The parameter r satisfying (21) can be considered as the rate of a variable length code. Note that if $r = 1$, then (21) coincides with the Kraft inequality.

6. AN UPPER BOUND ON THE DECODING ERROR PROBABILITY FOR VARIABLE LENGTH CODES

We will investigate the expected value of the probability $\Lambda_n^{(K)}(\widetilde{f}_{\rm sc},\widetilde{\varphi}_{\rm sc}|\ell)$ defined in (14), where the averaging is taken over following the code ensemble :

– *at each time instant, all symbols of the codewords* $\boldsymbol{x}(1),\ldots,\boldsymbol{x}(M)$ *are independently chosen from* $\mathcal{X}$ *in accordance with the probability distribution* $\{\,\omega(x), x\in\mathcal{X}\,\}$.

The probability $\Lambda_n^{(K)}(\widetilde{f}_{\rm sc},\widetilde{\varphi}_{\rm sc}|\ell)$ for at least one code does not exceed the expected value, and our results allow one to evaluate the potential characteristics of the encoding and decoding algorithms.

Notice that the definition of the code ensemble given above implies the independence of the code symbols belonging to the same codeword at each time instant t. Thus, the ensemble can be viewed as "completely randomized". However, methods of the analysis presented below can be also extended to the case when there are some restrictions on the structure of the code ensemble (for example, we can consider the ensemble of tree codes having a given topology and assign independent code symbols to all edges).

Lemma. *If* ℓ *is the length distribution of a variable length code and the parameter* r *satisfies* (21), *then*

$$\mathbb{E}\left[\,\lambda_n^{(k)}(\widetilde{f}_{\rm sc},\widetilde{\varphi}^*_{\rm sc}|\boldsymbol{m},\ell)\,\right]\le\Gamma(l_{\min},r)2^{-\ell(m_k)E_{\rm ran}(r)},\tag{23}$$

for any sequence of messages $\boldsymbol{m}=(m_1,m_2,\ldots)$, *where* $\mathbb{E}$ *denotes the averaging over the code ensemble defined above,* $\widetilde{\varphi}^*_{\rm sc}$ *denotes the maximum likelihood decoding,*

$$\Gamma(l_{\min},r)=\frac{1}{(1-2^{-l_{\min}E_{\rm ran}(r)})^2},\tag{24}$$

$l_{\min}$ *is the minimal codeword length defined in* (2), *and* $E_{\rm ran}(r)$ *is the value of the random coding bound defined in* (9).

Proof. Let us fix the sequence of messages $\boldsymbol{m}$ and choose the parameters τ and τ' in such a way that $\tau,\tau'\in\mathcal{T}(\boldsymbol{m})$ and $\tau\le t_k,\tau'>t_k$, where the set $\mathcal{T}(\boldsymbol{m})$ is defined in (11). Let $P_{\tau,\tau'}$ denote the probability that there exists a sequence $\widehat{\boldsymbol{x}}_{\tau'-\tau}$ of length $\tau'-\tau$ obtained as a concatenation of codewords corresponding to the messages $\widehat{m}_1,\ldots,\widehat{m}_j$ for some $j\ge 1$ such that

$$\mathcal{T}(\widehat{\boldsymbol{m}})\bigcap\mathcal{T}(\boldsymbol{m})\bigcap\{\tau+1,\ldots,\tau'-1\}=\varnothing$$

which is more likely to be transmitted than the sequence $(x_{\tau+1},\ldots,x_{\tau'})$ when $(y_{\tau+1},\ldots,\,y_{\tau'})$ is received at the output of the channel. Then, using (19), (22), and independence of the code symbols of the transmitted sequence and its competitor, we apply the random coding technique and obtain

$$\mathbb{E}\,[\,P_{\tau,\tau'}\,]\le\min_{0\le\rho\le 1}N^{\rho}_{\tau'-\tau}2^{-(\tau'-\tau)E_0(\rho)}\le 2^{-(\tau'-\tau)E_{\rm ran}(r)}.\tag{25}$$

The union bound on $\lambda_n^{(k)}(\widetilde{f}_{\rm sc},\widetilde{\varphi}^*_{\rm sc}|\boldsymbol{m},\ell)$ leads to the inequality

$$\lambda_n^{(k)}(\widetilde{f}_{\rm sc},\widetilde{\varphi}^*_{\rm sc}|\boldsymbol{m},\ell)\le\sum_{\substack{\tau\in\mathcal{T}\\ \tau\le t_k}}\sum_{\substack{\tau'\in\mathcal{T}\\ \tau>t_k}}P_{\tau,\tau'}\,,$$

and using (25) we have

$$\mathbb{E}\left[\,\lambda_n^{(k)}(\widetilde{f}_{\rm sc},\widetilde{\varphi}^*_{\rm sc}|\boldsymbol{m},\ell)\,\right]\le\sum_{\substack{\tau\in\mathcal{T}\\ \tau\le t_k}}\sum_{\substack{\tau'\in\mathcal{T}\\ \tau>t_k}}2^{-(\tau'-\tau)E_{\rm ran}(r)}.\tag{26}$$

However,

$$\tau = t_{k-i},\ \tau' = t_{k+i'} \implies \tau' - \tau = \sum_{j=k-i}^{k+i'} \ell(m_j) \geq \ell(m_k) + (i+i')l_{\min}$$

and if $E_{\mathrm{ran}}(r) > 0$, then using this statement in (26) we obtain

$$\mathbb{E}\left[\lambda_n^{(k)}(\widetilde{f}_{\mathrm{sc}}, \widetilde{\varphi}_{\mathrm{sc}}^* | \boldsymbol{m}, \ell)\right] \leq 2^{-\ell(m_k)E_{\mathrm{ran}}(r)} \sum_{i\geq 0}\sum_{i'\geq 0} 2^{-(i+i')l_{\min}E_{\mathrm{ran}}(r)} = \Gamma(l_{\min}, r)2^{-\ell(m_k)E_{\mathrm{ran}}(r)}.$$

△

Note that the upper bound (23) does not depend on the probability distribution on the messages and can be also used when the distribution is unknown. This result can be interpreted in the following way : given a sequence of messages, one can derive an upper bound on the decoding error probability that essentially depends on lengths of the codewords corresponding to these messages, while all other lengths enter the bound only through the parameters r and $l_{\min}$. Note also that, for specific sequences, the coefficient $\Gamma(l_{\min}, r)$ can be decreased as compared to (24).

We will use the well–known facts [9] concerning the types of sequences whose components belong to the set $\mathcal{U}$ to analyze the exponent of the upper bound on the average decoding error probability that can be derived using (23). Namely, let $\mathcal{P}^{(n)}$ denote the set of all possible types in $\mathcal{U}^n$, i.e., $\mathcal{P}^{(n)}$ consists of all probability distributions $\widehat{P} = \{\widehat{P}(u), u \in \mathcal{U}\}$ such that $n\widehat{P}(u), u \in \mathcal{U}$, are integers. Let

$$T^n_{\widehat{P}} = \Big\{ m \in \mathcal{U}^n : \sum_{i=1}^n \chi\{m^{(i)} = u\} = n\widehat{P}(u), \text{ for all } u \in \mathcal{U} \Big\}.$$

The messages belonging to the set $T^n_{\widehat{P}}$ are understood as messages of type $\widehat{P} \in \mathcal{P}^{(n)}$. Let

$$\mathcal{U}_0^n = \bigcup_{\widehat{P}\in\mathcal{P}_0^{(n)}} T^n_{\widehat{P}}, \tag{27}$$

where

$$\mathcal{P}_0^{(n)} = \Big\{ \widehat{P} \in \mathcal{P}^{(n)} : H(\widehat{P}) \leq H(P) - \delta_n \Big\} \tag{28}$$

and $\delta_n > 0$ is fixed. Then

$$\frac{\log|\mathcal{P}^{(n)}|}{n} \to 0, \quad \text{as} \quad n \to \infty,$$

$$P_m = \exp_2\Big\{ n \sum_{u\in\mathcal{U}} \widehat{P}(u)\log P(u) \Big\}, \quad \text{for all} \quad m \in T^n_{\widehat{P}},$$

$$\log|T^n_{\widehat{P}}| \leq nH(\widehat{P}),$$

$$|\mathcal{U}_0^n| \leq |\mathcal{P}^{(n)}| \exp_2\{nH(P)\},$$

where $\exp_2 z = 2^z$ for all z. Furthermore, one can assign the sequence $\{\delta_n\}$ in such a way that

$$\delta_n \to 0,\ \delta_n\sqrt{n} \to \infty,\ \alpha_n = P(\mathcal{U}_0^n) \to 0, \quad \text{as} \quad n \to \infty. \tag{29}$$

Theorem 2 (the statement and basic ideas of the proof were presented in [10]). *There exist a vector of codeword lengths ℓ and a pair of functions $(\widetilde{f}_{\mathrm{sc}}, \widetilde{\varphi}_{\mathrm{sc}})$, belonging to the maximization domain in* (15) *that lead to the inequality*

$$\widetilde{e}_{\mathrm{sc}}^{(K)} \geq E_{\mathrm{ran}}(H(P)), \tag{30}$$

where E_{ran} is the exponent of random coding bound defined in (9).

Corollary. *If P is such that the value of the random coding bound coincides with the value of the sphere packing bound at $H(P)$, then the lower bound* (30) *is tight, i.e.,*

$$E_{\text{ran}}(H(P)) = E_{\text{sp}}(H(P)) \implies \widetilde{e}_{\text{sc}}^{(K)} = E_{\text{ran}}(H(P)).$$

Proof. Let us fix $\delta_n > 0$ and denote

$$l_0 = \log|\mathcal{U}_0^n| - \log\alpha_n, \tag{31}$$

where the set $\mathcal{U}_0^n$ and the probability α_n are defined in (27) and (29), respectively. Let us also fix the parameter $r < 1$ that will be specified later and define the length distribution of a code by

$$\ell(m) = \frac{1}{r}\begin{cases} l_0, & \text{if } m \in \mathcal{U}_0^n, \\ -\log P_m, & \text{if } m \notin \mathcal{U}_0^n, \end{cases} \tag{32}$$

where we omit rounding to integers since it does not affect asymptotic characteristics. Then inequality (21) is satisfied since

$$\sum_{m\in\mathcal{U}^n} 2^{-r\ell(m)} = 2^{-l_0}|\mathcal{U}_0^n| + \sum_{m\in\mathcal{U}^n} P_m = \alpha_n + 1 - \alpha_n = 1.$$

Furthermore,

$$\overline{l} = \frac{1}{r}\Big[\alpha_n l_0 - \sum_{m\notin\mathcal{U}_0^n} P_m \log P_m\Big] \le n\frac{H(P)+\beta_n}{r}, \tag{33}$$

where

$$\beta_n = \alpha_n\frac{\log|\mathcal{U}_0^n|}{n} - \frac{\alpha_n\log\alpha_n}{n} \le \alpha_n\Big(H(P) + \frac{\log|\mathcal{P}^{(n)}|}{n}\Big) + \frac{1}{2n}$$

Let us denote $c(r) = E_{\text{ran}}(r)/r$. Then, by (23) and (32), we write

$$\mathbb{E}\left[\Lambda_n^{(K)}(\widetilde{f}_{\text{sc}}, \widetilde{\varphi}_{\text{sc}}|\ell)\right] \le 2\Gamma(l_{\min}, r)\max\Big\{\pi_0(r), \pi_0'(r)\Big\}, \tag{34}$$

where

$$\pi_0(r) = \alpha_n 2^{-l_0 c(r)}, \quad \pi_0'(r) = \sum_{m\notin\mathcal{U}_0^n} P_m^{1+c(r)},$$

and simple transformations, which are based on (31) and properties of types, lead to the upper bounds

$$\begin{aligned}
\pi_0(r) &\le \alpha_n^{1+c(r)}|\mathcal{P}^{(n)}|^{-c(r)}\exp_2\Big\{-nc(r)H(P)\Big\}\\
\pi_0'(r) &\le |\mathcal{P}^{(n)}|\exp_2\Big\{-n\min_{\widehat{P}\notin\mathcal{P}_0^{(n)}}\Big[D(\widehat{P}\parallel P) - c(r)\sum_{u\in\mathcal{U}}\widehat{P}(u)\log P(u)\Big]\Big\}\\
&\le |\mathcal{P}^{(n)}|\exp_2\Big\{-nc(r)\min_{\widehat{P}\notin\mathcal{P}_0^{(n)}}\Big[-\sum_{u\in\mathcal{U}}\widehat{P}(u)\log P(u)\Big]\Big\}\\
&\le |\mathcal{P}^{(n)}|\exp_2\Big\{-nc(r)\min_{\widehat{P}\notin\mathcal{P}_0^{(n)}}\Big[-\sum_{u\in\mathcal{U}}\widehat{P}(u)\log\widehat{P}(u)\Big]\Big\}\\
&\le |\mathcal{P}^{(n)}|\exp_2\Big\{-nc(r)(H(P)-\delta_n)\Big\}.
\end{aligned}$$

Let us assign $r = H(P)+\beta_n$ and choose the sequence $\{\delta_n\}$ in such a way that (29) is satisfied. Then, due to (33), $\overline{l} \le n$, and the codes whose codewords have lengths defined by (31), (32) belong to the maximization domain in (15). Since $\beta_n \to 0$, exponents of $\pi_0(r)$ and $\pi_0'(r)$ tend to $E_{\rm ran}(H(P))$, as $n \to \infty$, and (30) follows from (34).

As it is well–known [9], the probability that the source generates a message having type $\widehat{P}$ is equal to $\exp_2\{-n(D(\widehat{P} \parallel P) + \delta_n^*)\}$, where $D(\widehat{P} \parallel P)$ is the divergence between the distributions $\widehat{P}$ and P (see (8)), while $\delta_n^* \to 0$ as $n \to \infty$. All messages of type $\widehat{P}$ are equally likely, and their number asymptotically coincides with $\exp_2\{nH(\widehat{P})\}$. Therefore, the decoding error probability for the subcode, which is used to encode these messages, is not less than $\exp_2\{-nE_{\rm sp}(H(\widehat{P}))\}$, where $E_{\rm sp}$ is defined in (10). As the probability $\Lambda_n^{(k)}(\widetilde{f}_{\rm sc}, \widetilde{\varphi}_{\rm sc}^*|\ell)$ is bounded from below by the decoding error probability for the messages of type $\widehat{P}$, considerations above imply

$$\frac{-\log \Lambda_n^{(k)}(\widetilde{f}_{\rm sc}, \widetilde{\varphi}_{\rm sc}^*|\ell)}{n} \leq D(\widehat{P} \parallel P) + E_{\rm sp}(H(\widehat{P})) + \delta_n^*.$$

If $\widehat{P}$ is chosen in such a way that (29) is satisfied, the divergence $D(\widehat{P} \parallel P)$ tends to zero, and $E_{\rm sp}(H(\widehat{P}))$ tends to $E_{\rm sp}(H(P))$, as $n \to \infty$, i.e.,

$$\widetilde{e}_{\rm sc}^{(K)} \leq E_{\rm sp}(H(P)),$$

which proves the corollary. $\triangle$

REFERENCES

1. Jelinek, F., *Probabilistic Information Theory,* New York: McGraw Hill, 1968.

2. Csiszár, I., Joint Source–Channel Error Exponent, *Probl. Control Inf. Theory,* 1980, vol. 9, no. 5, pp. 315–328.

3. Csiszár, I., On the Error Exponent of Source–Channel Transmission with a Distortion Threshold, *IEEE Trans. Inf. Theory,* 1982, vol. 28. no. 6. pp. 823–828.

4. Ahlswede, R., and Gács, P., Two Contributions to Information Theory, *Colloquia Mathematica Societatis János Bolai,* Keszthely, 1975.

5. Viterbi, A.J., Error Bounds for Convolutional Codes and Asymptotically Optimum Decoding Algorithm, *IEEE Trans. Inf. Theory,* 1967, vol. 13, no. 2, pp. 260–269.

6. Gallager, R.G., A Simple Derivation of the Coding Theorem and Some Applications, *IEEE Trans. Inf. Theory,* 1965, vol. 11, no. 1, pp. 3–18.

7. Viterbi, A.J., and Omura, J.K., *Principles of Digital Communication and Coding,* New York: McGraw Hill, 1979.

8. Balakirsky, V.B., Prefix Codes for One Search Problem, Proc. of the All–Union Conference *Reliability and Effectivness of Information Transmission Systems,* Odessa, 1986, pp. 58–59.

9. Csiszár, I., and Körner, J., *Information Theory: Coding Theorems for Discrete Memoryless Systems,* New York: Academic, 1981.

10. Balakirsky, V.B., Joint Source–Channel Coding with Variable Length Codes, Proc. 1997 IEEE Int. Symp. on Information Theory, Ulm, 1997, p. 491.