Random-Coding Error Exponent of Variable-Length Codes with a Single-Bit Noiseless Feedback

Shai Ginzach
Rafael Advanced Defense Systems Ltd.
Haifa 3102102, Israel
Email: shaigi@rafael.co.il

Neri Merhav Igal Sason
Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200003, Israel
Email: {merhav,sason}@ee.technion.ac.il

***Abstract*—We study the random-coding error exponent function of variable–length codes in the presence of a noiseless feedback channel, which is allowed to be used merely for a single bit feedback per each transmitted message. In this study, we harness results and analysis techniques from the theory of sequential hypothesis testing, and combine them with modern distance enumeration methods which are used in the literature on error exponents. For this setup, sometimes referred to as stop-feedback, we derive an *exact* single-letter expression for the random-coding error exponent over the binary symmetric channel. For symmetric discrete memoryless channels, the exact error exponent at zero rate is obtained, and a lower bound is provided for any other positive rate below capacity.**

I. Introduction

We address the problem of stop-feedback codes, which are variable–length (VL) codes that require feedback of a single bit to stop the transmission once the decoder is ready to decode [1]. Stop-feedback is closely related to automatic repeat request (ARQ) codes, the latter being an even more restricted class of VL codes, where a single fixed-block length, non-feedback code is used repeatedly until the decoder produces a reliable estimate. While the reliability function of VL coding with unlimited feedback is known for decades [2], the question of the reliability functions for both ARQ and stop-feedback over a general discrete memoryless channel (DMC) is still open. The study of these reliability functions has started about fifty years ago. Weldon [3] first considered stop-feedback strategies for the binary symmetric channel (BSC) and he showed that the reliability function, $E(R)$, is lower bounded by $C-R$, where C is the channel capacity, and R is the coding rate.

In [4], Forney used Gallager's bounding method [5, Ch. 5] to bound the best achievable error exponents (i.e., reliability functions) of erasure and list decoders over DMCs. Both exponents stem from decoders that optimally trade off between error probability and erasure probability or between error probability and average list-size. He then proposed an achiebability scheme that uses a randomly generated fixed-block-length code and a decoder with an erasure option to define an ARQ code. For symmetric DMCs and very noisy channels (VNCs), Forney showed that [4, eq. (57)]

$$E(R) \geq E_{\text{r,f}}(R) \triangleq E_{\text{sp}}(R) + C - R, \;\; R_\infty \leq R \leq C \quad (1)$$

where $E_{\text{sp}}(R)$ is the sphere-packing error exponent [5, p. 157] and R_∞ is the rate at which the sphere-packing error exponent becomes infinite. In [6, Ch. 10], the same bound was obtained using the method of types, analyzing a decoder which generalizes the maximum mutual information (MMI) decoder for constant composition (CC) codes. Later on, Viterbi [7] used Gallager's bounding technique to prove that this bound is tight for both the Gaussian channel and VNCs, and Telatar [8] used the method of types for CC codes to show that $E_{\text{r,f}}(R)$ is tight for any DMC with ARQ codes at zero rate. In addition, [9] shows that for the Z-channel, zero error probability can be achieved with an ARQ code and hence $E_{\text{r,f}}(R)$ is, in general, not a tight lower bound for any DMC.

In this paper, we address the problem of random-coding error exponents with stop-feedback coding. In particular, the main contributions of this work are as follows:

1. Obtaining a lower bound on the random-coding error exponent for symmetric DMCs which is tight at zero rate. This generalization of Telatar's result [8] to the stop-feedback case also emphasizes the intuitive structure of this bound.
2. Deriving an exact expression for the random-coding error exponent of the BSC, which coincides with the lower bound for ARQ codes on the right-hand side (r.h.s.) of (1), and a lower bound on random-coding error exponent for symmetric DMCs.

In addition, we provide a simple proof of (1) which gives rise to an alternative expression for $E_{\text{r,f}}(R)$.

Following the line of research of insightful earlier works (e.g., [10], [11]), this work further illuminates the important interplay between information theory and sequential hypothesis testing. Specifically, we combine and modify two mathematical sets of tools. The first, inspired by a statistical-mechanical point of view on random code ensembles, uses certain distance enumerators in order to analyze random-coding exponents, and the second is taken from the theory of sequential multiple hypothesis testing.

S. Ginzach was with the department of electrical engineering at the Technion and this paper is part of his Master's thesis.

II. Notation Conventions

Throughout the paper, random variables (RVs) are denoted by capital letters, their realizations are denoted by the corresponding lower case letters, and their alphabets are denoted by calligraphic letters. An infinite sequence of RVs is denoted by bold face font. The length of finite sequences or, alternatively, the index at which an infinite sequence is truncated, appears as a superscript. Probability measures are denoted by P and Q, and their dimension is determined by the dimension of the sequence in the argument. For example, $P(\mathbf{x}^n)$ is the restriction of P to the σ-algebra $\mathcal{F}_n = \sigma(X_1, \ldots X_n)$. The indicator function of an event $\mathscr{A}$ is denoted by $\mathbb{I}\{\mathscr{A}\}$. We denote the binary entropy function by $h_2(p)$ and the binary relative entropy by $d_2(p\|q)$, where $p, q \in [0,1]$.

Define a random codebook $\{\mathbf{x}(0), \mathbf{x}(1) \ldots, \mathbf{x}(M-1)\}$, where $\mathbf{x}(i) = (x_1(i), x_2(i) \ldots)$ is the codeword assigned to the message i, chosen at random according to the measure P_X. Furthermore, let $\mathbf{z} = (z_1, z_2, \ldots)$ be the sequence

$$z_k \in \mathcal{X}^M \times \mathcal{Y},\ z_k = \{x_k(0), x_k(1), \ldots, x_k(M-1), y_k\}$$

where $\mathbf{y} = (y_1, y_2, \ldots)$ is the observed sequence at the output of the forward channel.

The forward channel is assumed to be a symmetric DMC, in the sense that the columns of the transition probability matrix $\left\{P_{Y|X}(y|x)\right\}$ are permutations of each other, which is the case, for example, with modulo–additive channels. For stop-feedback codes (as well as ARQ codes), only one feedback bit per message is allowed. Therefore, we assume that an instantaneous and error-free binary feedback channel is available.

Define the following M simple hypotheses:

$$\mathsf{H}_i : \Pr(\mathbf{z}^n) = P_i(\mathbf{z}^n), \quad i \in \{0, \ldots, M-1\}, \qquad (2)$$

where

$$P_i(\mathbf{z}^n) \triangleq P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}^n|\mathbf{x}^n(i)) \prod_{l=0}^{M-1} P_{\mathbf{X}}(\mathbf{x}^n(l)).$$

Note that for each i, $P_i(\mathbf{z}^n)$ is the distribution of the random process $\mathbf{z} = \{\mathbf{x}(0), \mathbf{x}(1) \ldots, \mathbf{x}(M-1), \mathbf{y}\}$, where all the $\mathbf{x}$'s are independent, and $\mathbf{y}$ is generated by sending $\mathbf{x}(i)$ through the DMC. In other words, if we assume that the M hypotheses are a-priori equiprobable, the problem of sequentially testing the hypotheses (2) is equivalent to deciding which one of the M sequences, $\mathbf{x}(0), \ldots, \mathbf{x}(M-1)$, was sent through the forward channel. Once a decision is made, the feedback channel is used to indicate that a new message should be sent. An important observation is that under hypothesis H_i, the random vectors $Z_1, Z_2, \ldots$ are i.i.d. We denote the class of sequential tests that select one of the hypotheses H_i in (2), by $\Delta = (N, d)$, where N denotes the stopping time, and d is the decision function.

The relative entropy from hypothesis H_i to hypothesis H_j is defined as

$$D(i\|j) \triangleq \mathbb{E}_i\left\{\log\left[\frac{P_i(Z)}{P_j(Z)}\right]\right\},$$

where $\mathbb{E}_i[\cdot]$ denotes the expected value under H_i. Let $D_i \triangleq \min_{j\neq i} D(i\|j)$. Note that the $D(i\|j) > 0$ since, by assumption, the probability measures of the M hypotheses are distinct. We shall also assume $D(i\|j) < \infty$. Symmetry considerations imply that D_i is independent of i, and hence will be denoted by D. Moreover, the inequality $C \leq D \leq C_1$ holds, where C is the capacity and C_1 is the reliability function at zero rate when no constraints are imposed on the feedback channel [2]. This result is intuitive when compared to the result in Sec. IV.

The block-length of stop-feedback codes is a function of the realization of the channel, and hence it is a random variable. It is customary to define the coding rate R and the reliability function $E(R)$ (i.e., the error exponent for optimal codes) in such cases to be

$$R = \frac{\log(M)}{\mathbb{E}[N]}, \quad E(R) = \limsup_{\mathbb{E}[N]\to\infty} \frac{-\log P_e}{\mathbb{E}[N]},$$

where P_e is the error probability. The error exponent for the average error probability with the optimal input distribution and random-coding is denoted by $E_{\mathrm{r}}(R)$.

III. Background and Known Results

Generalized versions of the results stated in this section can be found in the references herein.

A. Multi-hypothesis testing

Define M hypotheses $\mathsf{H}_i\colon P = P_i\ ,\ i \in \{0, \ldots, M-1\}$, where P_i are known distinct probability measures. Let $\mathbf{v}$ be an observation sequence, and denote the log-likelihood ratio processes with respect to (w.r.t.) a dominating measure Q by

$$L_i(n) = \log\left[\frac{P_i(v_1, \ldots, v_n)}{Q(v_1, \ldots, v_n)}\right], \quad i = 0, \ldots, M-1.$$

Let $W(j,i)$ be a given loss function associated with a decision on H_i when H_j is true, and let $(\pi_0, \pi_1 \ldots, \pi_{M-1})$ be the prior distribution vector of the hypotheses. We consider a Bayesian problem in which the risk, w.r.t. $W(j,i)$, associated with the average cost of deciding erroneously on H_i for $i \neq j$ is given by $R_i(\Delta) = \sum_{j=0, j\neq i}^{M-1} \pi_j W(j,i) P_j(d = i)$.

We introduce the following class of tests:

$$\mathbf{\Delta}(\rho) = \{\Delta : R_i(\Delta) \leq \rho_i\ ,\ i = 0, 1, \ldots, M-1\},$$

where $\rho = (\rho_0, \rho_1, \ldots, \rho_{M-1})$ is a given vector of positive finite numbers.

Next, we define the following stopping times:

$$N_i \triangleq \min_{n\geq 0}\left\{L_i(n) \geq a_i + \log\left(\sum_{j\neq i} w(j,i)\exp\big(L_j(n)\big)\right)\right\}, \qquad (3)$$

where $w(j,i) \triangleq \frac{\pi_j W(j,i)}{\pi_i}$, and $\{a_i\}$ are arbitrary positive thresholds. The test procedure $\Delta_a = (N_a, d_a)$ is defined as follows:

$$N_a = \min_{0\leq i\leq M-1} N_i, \quad d_a = i \text{ if } N_a = N_i. \qquad (4)$$

That is, we stop the data transmission as soon as the threshold in the r.h.s. of (3) is exceeded for some time index i and

decide in favor of the hypothesis H_i. This test is motivated by a Bayesian framework which was considered earlier, e.g., by Fishman [12], Golubev and Khas'minskii [13], and Baum and Veeravalli [14]. In this work, we take a particular interest in the case where $V_1, V_2, \ldots$ are i.i.d. under H_i. It will suffice to take $W(j,i)$ to be equal to 0 for $i=j$ and be equal to 1 otherwise, which is known as the $0-1$ loss function.

The following theorem concerns the asymptotic optimality of Δ_a for the i.i.d. case where each observation is, in general, a random vector.

Theorem 1 ([15]): Let N_a be the stopping rule of Δ_a. If the thresholds are chosen such that $a_i = \log \frac{\pi_i}{\rho_i}$ then, under the asymptotic regime of growing expected stopping times and a vanishing decision error probability

$$\inf_{\Delta \in \mathbf{\Delta}(\rho)} \mathbb{E}_i[N] = \mathbb{E}_i[N_a] = |\log \rho_i| / D_i \;\text{ as }\; \max_i \rho_i \to 0.$$

B. Erasure decoder and ARQ schemes

Consider the case of fixed block-length coding, where, at the end of transmission, the decoder has an additional option of not deciding, i.e., rejecting all messages. The resulting output is called an *erasure*. Under this setup, only if the decoder estimates the message incorrectly, we have an undetected error. It is clear that by allowing the erasure probability to increase, the undetected error probability can be reduced. A decoding scheme with an erasure option is a partition of the observation space $\mathcal{Y}^n$ into $(M+1)$ regions, $\mathcal{R}_0, \ldots, \mathcal{R}_M$. Such a decoder operates as follows: if the output sequence $\mathbf{y}^n \in \mathcal{Y}^n$ falls into $\mathcal{R}_i$ with $i \in \{0, \ldots, M-1\}$, then a decision is made in favor of message i. If $\mathbf{y}^n \in \mathcal{R}_M$ an erasure is declared. We will refer to the event $\{\mathbf{y}^n \in \mathcal{R}_M\}$ as the erasure event. Following Forney [4], we next define two additional undesired events. The event $\mathscr{E}_1$ corresponds to the case where the received vector does not fall in the decision region $\mathcal{R}_i$ of the transmitted message $i \in \{0, \ldots, M-1\}$. This event is the disjoint union of the erasure event and the event $\mathscr{E}_2$, which is the undetected error event, namely, the event of making the wrong decision. In [4], using the Neyman-Pearson theorem, Forney showed that the best trade-off between $\Pr(\mathscr{E}_1)$ and $\Pr(\mathscr{E}_2)$ is attained by the following decision regions:

$$\mathcal{R}_i^\star = \left\{ \mathbf{y}^n \in \mathcal{Y}^n : \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}^n|\mathbf{x}^n(i))}{\sum_{j \neq i} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}^n|\mathbf{x}^n(j))} \geq e^{nT} \right\}, \quad (5)$$

$$\mathcal{R}_M^\star = \bigcap_{i=0}^{M-1} (\mathcal{R}_i^\star)^c \quad (6)$$

where $T \geq 0$ is a free parameter which minimizes $\Pr(\mathscr{E}_1)$ for a given $\Pr(\mathscr{E}_2)$. Denote by $e_i(R,T)\,(i=1,2)$, the best achievable error exponents associated with $\Pr(\mathscr{E}_i)$, averaged over the ensemble of codes drawn i.i.d.; hence $e_i(R,T) \triangleq \limsup_{n\to\infty}\left[-\frac{1}{n}\log\left(\Pr(\mathscr{E}_i)\right)\right]$. In [4], lower bounds on the error exponents $e_1(R,T)$ and $e_2(R,T)$ were derived using Gallager's classical bounding method. In a more recent work [16], Somekh-Baruch and Merhav used distance enumerators in order to analyze random-coding exponents of an optimal decoder with an erasure option. Unlike the classical approach, their starting point was not a Gallager-type bound on the probability of error, based on the expectation of the sum of certain likelihood ratios, but rather the exact expression that defines the probability of an erasure and undetected errors. In other words, in [16], the authors derive exact single-letter expressions for the error exponents, in lieu of the lower bounds that were discussed so far. For example, for the BSC, $e_1(R,T)$ and $e_2(R,T)$ take on simple forms, as is apparent from the following theorem:

Theorem 2 ([16]): For the BSC with crossover probability $p < \frac{1}{2}$, let $\beta = \log\frac{1-p}{p}$. Under uniform random-coding, if $R \geq \log(2) - h_2\left(p + \frac{T}{\beta}\right)$, $e_1(R,T) = 0$ and otherwise

$$e_1(R,T) = \min\{d_2(\nu\|p) - h_2(\nu + T/\beta) + \log 2 - R\}$$

where the minimization is over $\nu \in \left[p, \delta_{\text{GV}}(R) - \frac{T}{\beta}\right]$ and $e_2(R,T) = e_1(R,T) + T$. The Gilbert-Varshamov (GV) distance, $\delta_{\text{GV}}(R)$, is defined to be the unique value $\delta \in [0, 1/2]$ for which $h_2(\delta) = \log 2 - R$.

For the forward and feedback channel models at hand, Forney proposed the following ARQ scheme [4]: the transmitter sends a codeword $\mathbf{x}^n(i) \in \mathcal{X}^n$, chosen at random from a codebook of rate R where $i \in \{0, \ldots, M-1\}$ and $M = \lceil e^{nR} \rceil$ is the total number of messages. After receiving a block of n symbols, the receiver uses an erasure-decoder, which decides that the transmitted codeword was $\mathbf{x}^n(i), i \in \{0, \ldots, M-1\}$, if and only if the received sequence $\mathbf{y}^n \in \mathcal{Y}^n$ falls in $\mathcal{R}_i^\star$, defined in (5). In this case, the receiver transmits an ACK message to the transmitter, and the transmitter sends the next message. If $\mathbf{y}^n \in \mathcal{R}_M^\star$ the receiver declares an erasure, and sends a NACK feedback bit. Upon receiving a NACK, the transmitter repeats the message. Note that in this scheme, the decoder discards the earlier received sequences, and it uses only the latest received n symbols for decoding. Using this ARQ scheme, (1) is shown to hold for symmetric DMCs and VNCs.

A new, short proof of this result can be obtained using Theorem 2 and the fact that, for the ARQ described above, the error exponent is lower bounded by $\lim_{e_1(R,T)\to 0} e_2(R,T)$. This enables us to focus only on the point in which $e_1(R,T)$ becomes positive. Specifically, for a BSC with crossover probability $p < \frac{1}{2}$, and an equiprobable input distribution for random-coding, it follows that

$$\lim_{e_1(R,T)\to 0} T = \beta\left[\delta_{\text{GV}}(R) - \delta_{\text{GV}}(C)\right] \equiv E_{\text{r,f}}(R). \quad (7)$$

This expression also gives rise to a simple alternative expression for $E_{\text{r,f}}(R)$ in terms of the difference between the GV distance at the capacity and the GV distance at rate R.

IV. Exact Error Exponent at Zero Rate

In this section, we consider $E_{\text{r}}(0)$ (i.e., when M is fixed or grows sub-exponentially with the expected value of the observation time), and a symmetric discrete forward channel. For this zero-rate regime, the fact that the problem of variable length coding with stop-feedback can be formalized as a sequential multiple-hypothesis testing problem, makes

it possible to apply Theorem 1. Specifically, the symmetry between the different hypotheses in the sequential hypothesis testing problem in Sec. III implies that $R_i(\Delta) = \frac{P_e(\Delta)}{M}$ for all $i = 0, \ldots, M-1$. Invoking Theorem 1 yields that for any fixed $\delta > 0$ and small enough ϵ,

$$\inf_{\Delta: P_e(\Delta) \le \epsilon} \mathbb{E}_i[N] \ge (1-\delta) \frac{1}{D} \log\left(\frac{M}{P_e(\Delta)}\right).$$

On the other hand, applying $\Delta_a = (N_a, d_a)$ to the hypothesis testing problem at hand yields

$$\mathbb{E}_i[N_a] \le (1+\delta) \frac{1}{D} \log\left(\frac{M}{P_e(\Delta)}\right).$$

Combining these results and using the definition for $E_{\mathrm{r}}(R)$ at $R = 0$, we get that $E_{\mathrm{r}}(0) = D = E_{\mathrm{r,f}}(0)$. This implies that for a large family of channels, including symmetric channels, under random-coding, $E_{\mathrm{r,f}}(0)$ is tight for stop-feedback coding. Moreover, it exemplifies the usefulness of formulating the communication problem at hand as a sequential hypothesis test.

V. Lower Bound on the Error Exponent Function

In this section, we obtain a lower bound on the error exponent function by applying Δ_a to (2). For the BSC this gives rise to an alternative proof for the achievability of $E_{\mathrm{r,f}}(R)$, and for symmetric DMCs, this analysis yields a new lower bound on the error exponent.

The main challenge in this section and in Section VI is that neither Theorem 1 nor other sequential hypothesis testing analysis tools lend themselves easily to the asymptotic regime in which M increases exponentially with the expected value of the observation time. In spite of this obstacle, we next show how the performance analysis of Δ_a can be modified to take into account an arbitrary $R > 0$.

Let $i \in \{0, \ldots, M-1\}$ and $a \ge 0$. Applying the stopping rule N_i, defined in (3), to (2) yields

$$N_i = \min_{n \ge 0} \{\Lambda_i(n) \ge a\}, \tag{8}$$

$$\Lambda_i(n) \triangleq \log\left[\frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}^n|\mathbf{x}^n(i))}{\sum_{j \ne i} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}^n|\mathbf{x}^n(j))}\right]. \tag{9}$$

Assume, without loss of generality, that the message corresponding to H_0 was sent. By definition of N_a in (4) and N_0 in (8), $N_a \le N_0$, and so $\mathbb{E}_0[N_a] \le \mathbb{E}_0[N_0]$. For an arbitrary $\bar{n} \in \mathbb{N}$, one can bound $P_0(N_0 \ge n)$ for all $n \le \bar{n}$, and get

$$\mathbb{E}_0[N_a] \le \bar{n} + \sum_{n > \bar{n}} P_0(N_0 \ge n). \tag{10}$$

By definition of N_0, each term of the sum on the right side of (10) can be bounded as $P_0(N_0 \ge n) \le P_0(\Lambda_0(n) < a)$, and the expression on the right side lends itself to analysis using distance enumerators, akin to [16] and [17]. This probability is also closely related to the probability that Forney's erasure decoder does not make the right decision when a random code of block length n is used as it is implied by (5) and (6). Note that, for Δ_a,

$$P_e(\Delta_a) = \mathbb{E}_0\left[\mathbb{I}\{N_a = N_0, N_0 < \infty\} \frac{\sum_{j=1}^{M-1} P_j(\mathbf{Z})}{P_0(\mathbf{Z})}\right]$$

and since $N_a = N_0$ implies $\Lambda_0(N_a) \ge a$ a.s., $P_e(\Delta_a) \le e^{-a}$.

Define $R_n \triangleq \frac{\log M}{n}$ and $\bar{n} \triangleq \max_{n \in \mathbb{N}} \{e_1(R_n, T) \ge 0\}$. For example, for the BSC, using Theorem 2 yields

$$\bar{n} = \max_{n \in \mathbb{N}} \left\{R_n \ge \log 2 - h_2\left(p + \frac{a}{\beta n}\right)\right\}. \tag{11}$$

Note that $\bar{n} \to \infty$ as $M \to \infty$ and $\bar{n} \le \frac{a}{E_{\mathrm{r,f}}(R_{\bar{n}})}$, which implies that $E_{\mathrm{r}}(R) \ge E_{\mathrm{r,f}}(R)$. A proof that $E_{\mathrm{r,f}}(R)$ is achievable was given in [4] using ARQ coding (see also (7)). In [18], Gopola *et al.* commented that this error exponent is not improved even with incremental redundancy coding, that is, without discarding the blocks that produced a NACK message at the receiver. In Section VI we show that $E_{\mathrm{r,f}}(R)$ is, in fact, the best error exponent that can be achieved using random coding.

Next, we generalize this result for DMCs with the symmetry property under which for every real s, $\gamma_y(s) = \sum_{x \in \mathcal{X}} P(x) P^s(y|x)$ is independent of y, which is the case for symmetric channels in the sense defined above. In this case $\gamma_y(s)$ will be denoted by $\gamma(s)$. Let s_R be the solution to the equation $\gamma(s) - s\gamma'(s) = R$. Analogously to the use of Theorem 2 in defining $\bar{n}$ in (11), we use Corollary 1 in [16] (which is not specified here due to space limitations) in order to define $\bar{n}$ for this class of channels:

$$\bar{n} = \bar{n}(a, M) \triangleq \max_{n \in \mathbb{N}} \{\gamma'(s_{R_n}) - H(Y|X) \le a/n\}.$$

Analysis of (10) yields

$$E_{\mathrm{r}}(R) \ge \gamma'(s_R) - H(Y|X),$$

where $H(Y|X)$ is the conditional entropy of Y given X.

VI. Upper Bound for the BSC

Let $\Delta = (N, d)$ be an optimal sequential multiple hypothesis test for (2) over a BSC with crossover probability p in the error exponent sense, and let $P_e = P_e(\Delta)$ be the associated decoding error probability. Define the event $\mathscr{A}_{i,\bar{n}} \triangleq \{d = i, N \le \bar{n}\}$, and note that for any a and $\bar{n}$,

$$\begin{aligned}
\sum_{j \ne i} P_j(d = i) &= \sum_{j \ne i} \sum_{\mathbf{z}} \mathbb{I}\{d = i\} P_j(\mathbf{z}) \\
&= \sum_{\mathbf{z}} \sum_{j \ne i} \mathbb{I}\{d = i\} \frac{P_j(\mathbf{z})}{P_i(\mathbf{z})} P_i(\mathbf{z}) \\
&= \mathbb{E}_i\left[\mathbb{I}\{d = i\} \frac{\sum_{j \ne i} P_j(\mathbf{Z})}{P_i(\mathbf{Z})}\right] \\
&\ge \mathbb{E}_i\left[\mathbb{I}\{\mathscr{A}_{i,\bar{n}}, \Lambda_i(N) < a\} e^{-\Lambda_i(N)}\right] \\
&\ge e^{-a} P_i\left(\mathscr{A}_{i,\bar{n}}, \sup_{n \le \bar{n}} \Lambda_i(n) < a\right).
\end{aligned} \tag{12}$$

Using the union bound in (12), the Markov inequality on $P_i\left(N>\bar{n}\right)$ and some algebra, it follows that

$$\frac{\mathbb{E}\left[N\right]}{\bar{n}} \geq 1-P_e\left(e^a+1\right)-P_i\left(\sup_{n\leq\bar{n}}\Lambda_i\left(n\right)>a\right). \quad (13)$$

In order to further bound the r.h.s. of (13) we use the following lemma:

Lemma 3: Let $\mathcal{F}_n$ be the filtration generated by $\mathbf{Z}^n$. For symmetric DMCs, $\left(\Lambda_i\left(n\right),\mathcal{F}_n\right)$ is a submartingale w.r.t. P_i. To show that $\Lambda_n\in\mathcal{F}_n$ and $\mathbb{E}_i\left[\left|\Lambda_i\left(n\right)\right|\right]<\infty$ is straightforward, so in order to prove the claim it is left to show that for all n

$$\Lambda_i\left(n\right)\leq\mathbb{E}_i\left[\Lambda_i\left(n+1\right)|\mathcal{F}_n\right]. \quad (14)$$

Towards that end, note that the r.h.s. of (14) is equal to

$$\log\left[P_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{Y}^n|\mathbf{X}^n\left(i\right)\right)\right]-H\left(Y|X\right)-\mathbb{E}_i\left\{\log\left[\sum_{j=0,j\neq i}^{M-1}P_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{Y}^{n+1}|\mathbf{X}^{n+1}\left(j\right)\right)\right]\middle|\mathcal{F}_n\right\},$$

and by Jensen's inequality,

$$\mathbb{E}_i\left\{\log\left[\sum_{j=0,j\neq i}^{M-1}P_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{Y}^{n+1}|\mathbf{X}^{n+1}\left(j\right)\right)\right]\middle|\mathcal{F}_n\right\} \leq\log\left[\sum_{j=0,j\neq i}^{M-1}P_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{Y}^{n}|\mathbf{X}^{n}\left(j\right)\right)\right]-H\left(Y|X\right).$$

Combining these results with (9) yields (14).

Let ϵ and μ be arbitrarily small positive numbers, and let $a\triangleq-\left(1-\epsilon\right)\log P_e+\mu$ and $\bar{n}\triangleq\left(1+\epsilon\right)\mathbb{E}\left[N\right]$. Note that, by Lemma 3, $\Lambda_i\left(n\right)+\left(1-\epsilon\right)\log P_e$ is also a submartingale. Applying Doob's inequality [19, Theorem 5.4.2] yields

$$P_i\left(\sup_{n\leq\bar{n}}\Lambda_i\left(n\right)>a\right)\leq\frac{\mathbb{E}_i[\Lambda_i\left(\bar{n}\right)+\left(1-\epsilon\right)\log P_e]^+}{\mu}, \quad (15)$$

where $\left[x\right]^+=\max\left\{x,0\right\}$. Henceforth, we consider only the BSC case. Define $\bar{R}\triangleq\frac{R}{1+\epsilon}$ and note that

$$\Lambda_i\left(\bar{n}\right)=\log\left[\frac{e^{-\bar{n}\beta\delta_i(\bar{n})}}{\sum_\delta N_{\mathbf{y}^{\bar{n}}}\left(\bar{n}\delta\right)e^{-\bar{n}\beta\delta}}\right], \quad (16)$$

where $N_{\mathbf{y}^{\bar{n}}}\left(\bar{n}\delta\right)$ is the number of incorrect codewords whose Hamming distance from $\mathbf{y}^{\bar{n}}$ is $\bar{n}\delta$, and $\bar{n}\delta_i\left(\bar{n}\right)$ is the Hamming distance between $\mathbf{y}^{\bar{n}}$ and $\mathbf{x}^{\bar{n}}\left(i\right)$. In order to further bound the r.h.s. of (15) we use a result from [17, Ch. 6] that states that with high probability (double-exponentially with $\bar{n}$),

$$\left\{e^{\bar{n}\left[\bar{R}+h_2(\delta)-\log(2)-\epsilon\right]}\leq N_{\mathbf{y}^{\bar{n}}}\left(\bar{n}\delta\right)\leq e^{\bar{n}\left[\bar{R}+h_2(\delta)-\log(2)+\epsilon\right]}\right\},$$

for any $\delta\in\mathcal{G}_{\bar{R}}\triangleq\left\{\delta\in[0,1]:\left[\delta_{\text{GV}}\left(\bar{R}\right),1-\delta_{\text{GV}}\left(\bar{R}\right)\right]\right\}$. The event $\left\{\left|\delta_i\left(\bar{n}\right)-p\right|\leq\epsilon\right\}$ also holds with high probability (exponentially with $\bar{n}$). An important observation is that under the intersection of these events,

$$-\bar{n}\beta\delta_i\left(\bar{n}\right)-\log\left(N_{\mathbf{y}^{\bar{n}}}\left(\bar{n}\bar{\delta}\right)e^{-\bar{n}\beta\bar{\delta}}\right)\leq\bar{n}\left[E_{\text{r,f}}\left(\bar{R}\right)+\epsilon\left(1+\beta\right)\right].$$

for any $\bar{\delta}\in\mathcal{G}_{\bar{R}}$, and hence $\Lambda_i\left(\bar{n}\right)$ is bounded by $\bar{n}\left[E_{\text{r,f}}\left(\bar{R}\right)-\epsilon\left(1+\beta\right)\right]$. Under the complementary of this intersection, $\Lambda_i\left(\bar{n}\right)$ is trivially bounded by a polynomial function of $\bar{n}$, but the probability of this event decays exponentially. We use these results in order to bound the expected value in (15). On substituting in a and $\bar{n}$ and taking $\mathbb{E}\left[N\right]\to\infty$ and $P_e\to0$, we conclude that

$$\lim\left\{\bar{n}\left[E_{\text{r,f}}\left(\bar{R}\right)+\epsilon\left(1+\beta\right)\right]+\left(1-\epsilon\right)\log P_e\right\}\geq0$$

and hence $E_{\text{r}}\left(R\right)\leq E_{\text{r,f}}\left(R\right)$. In Sec. V we showed that $E_{\text{r}}\left(R\right)\geq E_{\text{r,f}}\left(R\right)$ which implies the following theorem:

Theorem 4: For stop-feedback over a BSC, the best achievable random-coding error exponent is $E_{\text{r,f}}\left(R\right)$, as given in (7).

REFERENCES

[1] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4903–4925, 2011.

[2] M. V. Burnashev, "Data transmission over a discrete channel with feedback. random transmission time," *Problemy peredachi informatsii*, vol. 12, no. 4, pp. 10–30, 1976.

[3] E. J. Weldon, "Asymptotic error coding bounds for the binary symmetric channel with feedback," *University of Florida*, 1963.

[4] G. Forney Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Transactions on Information Theory*, vol. 14, no. 2, pp. 206–220, March 1968.

[5] R. G. Gallager, "Information theory and reliable communication," 1968.

[6] I. Csiszar and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[7] A. Viterbi, "Error bounds for the white Gaussian and other very noisy memoryless channels with generalized decision regions," *Information Theory, IEEE Transactions on*, vol. 15, no. 2, pp. 279–287, March 1969.

[8] I. E. Telatar, "Multi-access communications with decision feedback decoding," Ph.D. dissertation, Massachusetts Institute of Technology, 1992.

[9] I. E. Telatar, L. Louis, and R. Gallager, "New exponential upper bounds to error and erasure probabilities," in *Information Theory, 1994. Proceedings., 1994 IEEE International Symposium on*. IEEE, June 1994, p. 379.

[10] P. Berlin, B. Nakiboglu, B. Rimoldi, and E. Telatar, "A simple converse of burnashev's reliability function," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3074–3080, 2009.

[11] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," *The Annals of Statistics*, vol. 41, no. 6, pp. 2703–2738, 2013.

[12] M. M. Fishman, "Average duration of asymptotically optimal multialternative sequential procedure for recognition of processes," *Soviet Journ. Communic. Technol. Electron*, vol. 30, pp. 2541–2548, 1987.

[13] G. K. Golubev and R. Z. Khas'minskii, "Sequential testing for several signals in Gaussian white noise," *Theory of Probability & Its Applications*, vol. 28, no. 3, pp. 573–584, 1984.

[14] C. W. Baum and V. V. Veeravalli, "A sequential procedure for multihypothesis testing," *IEEE Transactions on Information Theory*, vol. 40, no. 6, November 1994.

[15] V. P. Draglia, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests. I. asymptotic optimality," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2448–2461, November 1999.

[16] A. Somekh-Baruch and N. Merhav, "Exact random coding exponents for erasure decoding," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6444–6454, October 2011.

[17] N. Merhav, "Statistical physics and information theory," *Foundation and Trends in communications and Information Theory*, vol. 6, no. 1–2, 2010.

[18] P. K. Gopala, Y. Nam, and H. El Gamal, "On the error exponents of ARQ channels with deadlines," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 4265–4273, November 2007.

[19] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2010, vol. 3.