Feedback in the non-asymptotic regime

Yury Polyanskiy, H. Vincent Poor, and Sergio Verdú

Abstract

Without feedback, the backoff from capacity due to non-asymptotic blocklength can be quite substantial for blocklengths and error probabilities of interest in many practical applications. In this paper, novel achievability bounds are used to demonstrate that in the non-asymptotic regime, the maximal achievable rate improves dramatically thanks to variable-length coding and feedback. For example, for the binary symmetric channel with capacity $1/2$ the blocklength required to achieve 90% of the capacity is smaller than 200, compared to at least 3100 for the best fixed-blocklength code (even with noiseless feedback).

Virtually all the advantages of noiseless feedback are shown to be achievable even if the feedback link is used only to send a single signal informing the encoder to terminate the transmission (stop-feedback). It is demonstrated that the non-asymptotic behavior of the fundamental limit depends crucially on the particular model chosen for the "end-of-packet" control signal. Fixed-blocklength codes and related questions concerning communicating with a guaranteed delay are discussed, in which situation the feedback is demonstrated to be almost useless even non-asymptotically.

Index Terms

Shannon theory, channel capacity, feedback, stop-feedback, non-asymptotic analysis, memoryless channels, achievability bounds, converse bounds.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544 USA. e-mail: {ypolyans,poor,verdu}@princeton.edu.

This work was supported by the National Science Foundation under Grants CCF-06-35154, CCF-10-16625 and CNS-09-05398.

I. INTRODUCTION

In the context of fixed blocklength communication, Shannon [1] showed that noiseless feedback does not increase the capacity of memoryless channels but can increase the zero-error capacity. For a class of symmetric discrete memoryless channels (DMCs), Dobrushin [2] demonstrated that the sphere-packing bound holds even in the presence of noiseless feedback.

Nevertheless, it is known that feedback can be very useful provided that variable-length codes are allowed. In his ground-breaking contribution, Burnashev [3] demonstrated that the error exponent improves in this setting and admits a particularly simple expression:

$$E(R) = \frac{C_1}{C}(C - R)\,, \tag{1}$$

for all rates $0 < R < C$, where C is the capacity of the channel and C_1 is the maximal relative entropy between the conditional output distributions. Moreover, zero-error capacity may improve from zero to the Shannon capacity (as in the case of the binary erasure channel (BEC)) if variable length is allowed. Furthermore, since existing communication systems with feedback (such as ARQ) have variable length, in the analysis of fundamental limits for channels with feedback, it is much more relevant and interesting to allow codes whose length is allowed to depend on the channel behavior.

We mention a few extensions of Burnashev's work [3], [4] relevant to this paper. Yamamoto and Itoh proposed a simple and conceptually important two-phase coding scheme, attaining the optimal error exponent [5]. Using the notion of Goppa's empirical mutual information (EMI) several authors have constructed universal coding schemes attaining rates arbitrarily close to capacity with small probability of error [6], [7], exponentially decaying probability of error [8] and even attaining the optimal Burnashev exponent [9], [10] simultaneously for a collection of channels. An extension to arbitrary varying channels with full state information available at the decoder has been recently proposed as well [11].

In contrast to the error exponent analysis of variable-length coding with feedback, which focuses on the regime of asymptotically long average blocklength at fixed rate, in this paper, following [12] we focus on the regime of fixed probability of error and finite average blocklength. Another aspect that was not previously addressed in the literature is the following. In practice, control information (such as initiation and termination) is not under the purview of the physical layer. However, the information theory literature typically assumes that all the feed-forward

control information is carried through the same noisy channel as the information payload. This is most notably illustrated by Burnashev's model in which the error exponent is, in fact, dictated by the reliability with which the termination information is conveyed to the receiver through the DMC while at the same time assuming that the feedback link has infinite reliability to carry not just a termination symbol but the whole sequence of channel outputs. To separate physical-channel issues from upper-layer issues, and avoid mismodelling of control signaling, it is important to realize that initiation/termination symbols are in fact carried through layers and protocols whose reliabilities need not be similar to those experienced by the payload. To capture this, we propose a simple modification of the (forward) channel model through the introduction of a "use-once" termination symbol whose transmission disables further communication.

The organization of this paper is as follows. Section II presents a formal statement of the problem and examines the relationships between different definitions of variable-length coding. Section III analyzes the maximal achievable rate with and without a termination symbol. Section IV focuses on zero-error communication. Section V discusses fixed-blocklength coding with feedback and problems related to transmitting with guaranteed delay, arising in communication systems with real-time data.

II. Statement of the problem

In this paper we consider the following channel coding scenario. A non-anticipatory channel consists of a pair of input and output alphabets $\mathcal{A}$ and $\mathcal{B}$ together with a sequence of conditional probability kernels $\{P_{Y_i|X_1^iY_1^{i-1}}\}_{i=1}^{\infty}$. Such channel is called (stationary) memoryless if

$$P_{Y_i|X_1^iY_1^{i-1}} = P_{Y_i|X_i} = P_{Y_1|X_1}\,, \quad \forall i \geq 1 \tag{2}$$

and if $\mathcal{A}$ and $\mathcal{B}$ are finite, it is known as a DMC.

Definition 1: An (ℓ, M, ϵ) variable-length feedback (VLF) code, where ℓ is a positive real, M is a positive integer and $0 \leq \epsilon \leq 1$, is defined by:

1) A space $\mathcal{U}$ with[1] $|\mathcal{U}| \leq 3$ and a probability distribution P_U on it, defining a random variable U which is revealed to both transmitter and receiver before the start of transmission; i.e. U acts as common randomness used to initialize the encoder and the decoder before the start of transmission.

[1]The bound on the cardinality of $\mathcal{U}$ is justified by Theorem 19 in the appendix.

2) A sequence of encoders $f_n : \mathcal{U} \times \{1, \ldots, M\} \times \mathcal{B}^{n-1} \to \mathcal{A}$, $n \geq 1$, defining channel inputs

$$X_n = f_n(U, W, Y^{n-1}) \,, \tag{3}$$

where $W \in \{1, \ldots, M\}$ is the equiprobable message.

3) A sequence of decoders $g_n : \mathcal{U} \times \mathcal{B}^n \to \{1, \ldots, M\}$ providing the best estimate of W at time n.

4) A non-negative integer-valued random variable τ, a stopping time of the filtration $\mathscr{G}_n = \sigma\{U, Y_1, \ldots, Y_n\}$, which satisfies

$$\mathbb{E}\,[\tau] \;\; \leq \;\; \ell \,. \tag{4}$$

The final decision $\hat{W}$ is computed at the time instant τ:

$$\hat{W} = g_\tau(U, Y^\tau) \,, \tag{5}$$

and must satisfy

$$\mathbb{P}[\hat{W} \neq W] \;\; \leq \;\; \epsilon \,. \tag{6}$$

The fundamental limit of channel coding with feedback is given by the following quantity:

$$M_f^*(\ell, \epsilon) = \max\{M : \exists(\ell, M, \epsilon)\text{-VLF code}\} \,. \tag{7}$$

Those codes that do not require the availability of U, i.e. the ones with $|\mathcal{U}| = 1$, are called *deterministic* codes. Although from a practical viewpoint there is hardly any motivation to allow for non-deterministic codes, they simplify the analysis and expressions, just like randomized tests do in hypothesis testing. Also similar to the latter, the difference in performance between the deterministic and non-deterministic codes is negligible for any practically interesting M and ℓ, since a few initial channel outputs can be used to supply any required common randomness.

In a VLF code the decision about stopping transmission is taken solely upon observation of channel outputs in a causal manner. This is the setup investigated by Burnashev [3]. Note that since τ is computed at the decoder, it is not necessary to specify the values of $g_n(Y^n)$ for $n \neq \tau$. In this way the decoder is a map $g : \mathcal{B}^\infty \to \{1, \ldots, M\}$ measurable with respect to $\mathscr{G}_\tau$.

Definition 2: An (ℓ, M, ϵ) variable-length feedback code with termination (VLFT), where ℓ is a positive real, M is a positive integer and $0 \leq \epsilon \leq 1$, is defined similarly to VLF codes with an exception that condition 4) in the Definition 1 is replaced by

4') A non-negative integer-valued random variable τ, a stopping time of the filtration $\mathscr{G}_n = \sigma\{W, U, Y_1, \ldots, Y_n\}$, which satisfies

$$\mathbb{E}[\tau] \quad \leq \quad \ell\,. \tag{8}$$

The fundamental limit of channel coding with feedback and termination is given by the following quantity:

$$M_{\mathsf{t}}^*(\ell, \epsilon) = \max\{M : \exists(\ell, M, \epsilon)\text{-VLFT code}\}\,. \tag{9}$$

In a VLFT code, "termination" is used to indicate the fact that the practical realization of such a coding scheme requires a method of sending a reliable end-of-packet signal by means other than using the $\mathcal{A} \to \mathcal{B}$ channel (e.g., by cutting off a carrier). As we discussed in the introduction, timing (including termination) is usually handled by a different layer in the protocol. Note that equivalently, a VLFT code may be understood as a VLF code used over a modified channel, having an additional special use-once input symbol, transmission of which disables further communication (see the proof of Theorem 4 below for a concrete application of this idea). We prefer, however, to understand the channel as a fixed stochastic model, while the structural constraints (such as how precisely the transmission terminates, or whether the feedback is available) are left to the definition of the code.

The following are examples of VLFT codes:

1) VLF codes are a special case in which the stopping time τ is determined autonomously by the decoder; due to availability of the feedback, τ is also known to the encoder so that transmission can be cut off at τ.
2) *stop-feedback codes* are a special case of VLF codes where the encoder functions $\{f_n\}_{n=1}^{\infty}$ satisfy:

$$f_n(U, W, Y^{n-1}) = f_n(U, W)\,. \tag{10}$$

   Such codes require very limited communication over feedback: only a single signal to stop the transmission once the decoder is ready to decode.
3) variable-length codes (without feedback), or *VL codes*, defined in [20, Problem 2.1.25] and [19], are VLFT codes required to satisfy two additional requirements: τ is a function of (W, U) and the encoder is not allowed to use feedback, i.e. (10) holds. The fundamental

limit and the ϵ-capacity of variable-length codes are given by

$$M_v^*(\ell,\epsilon) = \max\{M : \exists(\ell, M, \epsilon)\text{-VL code}\}\,, \tag{11}$$

$$[\![C_\epsilon]\!] = \lim_{\ell\to\infty}\frac{1}{\ell}\log M_v^*(\ell,\epsilon)\,. \tag{12}$$

4) fixed-to-variable codes, or *FV codes*, defined in [19] are also required to satisfy (10), while the stopping time is[2]

$$\tau = \inf\{n \geq 1 : g_n(U, Y^n) = W\}\,, \tag{13}$$

and therefore, such codes are zero-error VLFT codes. Of course, not all zero-error VLFT codes are FV codes, since in general condition (10) does not necessarily hold.

5) automatic repeat request (*ARQ*) codes analyzed in [12, Section IV.E] are yet a more restricted class of deterministic FV codes, where a single fixed-blocklength, non-feedback code is used repeatedly until the decoder produces a correct estimate.

The main goal of this paper is to analyze the behavior of $\log M_f^*(\ell,\epsilon)$ and $\log M_{\mathtt{t}}^*(\ell,\epsilon)$ and compare them with the behavior of the fundamental limit without feedback, $\log M^*(n,\epsilon)$. Regarding the behavior of $\log M_f^*(\ell,\epsilon)$ Burnashev's result (1) can be restated as

$$\log M_f^*(\ell, \exp\{-E\ell\}) = \ell C\left(1 - \frac{E}{C_1}\right) + o(\ell)\,, \tag{14}$$

for any $0 < E < C_1$. Although (14) does not imply any statement about the expansion of $\log M_f^*(\ell,\epsilon)$ for a fixed ϵ, it still demonstrates that in the regime of very small probability of error, the parameter C_1 emerges as an important quantity.

III. Fundamental limits for $\epsilon > 0$.

A. Main results

The first result shows that, under variable-length coding, allowing a non-vanishing error probability ϵ boosts the ϵ-capacity by a factor of $\frac{1}{1-\epsilon}$ even in the absence of feedback.

Theorem 1: For any non-anticipatory channel with capacity C that satisfies the strong converse for fixed-blocklength codes (without feedback), the ϵ-capacity under variable-length coding without feedback, cf. (12), is

$$[\![C_\epsilon]\!] = \frac{C}{1-\epsilon}\,, \epsilon \in (0,1)\,. \tag{15}$$

[2]As explained in [19], this model encompasses fountain codes in which the decoder can get a highly reliable estimate of τ autonomously without the need for a termination symbol.

The proof is given in the appendix. In general, it is known [19, Theorem 16] that the VL capacity, $[\![C]\!] = \lim_{\epsilon\to 0}[\![C_\epsilon]\!]$, is equal to the conventional fixed-blocklength capacity without feedback, C, for any non-anticipatory channel (not necessarily satisfying the strong converse). On the other hand, the capacity of FV codes for state-dependent non-ergodic channels can be larger than C [19].

Our main result is the following:

Theorem 2: For an arbitrary DMC with capacity C we have for any $0 < \epsilon < 1$

$$\log M_f^*(\ell,\epsilon) = \frac{\ell C}{1-\epsilon} + O(\log \ell)\,, \tag{16}$$

$$\log M_{\mathtt{t}}^*(\ell,\epsilon) = \frac{\ell C}{1-\epsilon} + O(\log \ell)\,. \tag{17}$$

More precisely, we have

$$\frac{\ell C}{1-\epsilon} - \log \ell + O(1) \le \log M_f^*(\ell,\epsilon) \le \frac{\ell C}{1-\epsilon} + O(1)\,, \tag{18}$$

$$\log M_f^*(\ell,\epsilon) \le \log M_{\mathtt{t}}^*(\ell,\epsilon) \le \frac{\ell C + \log \ell}{1-\epsilon} + O(1)\,. \tag{19}$$

A consequence of Theorem 2 is that for DMCs, feedback (even in the setup of VLFT codes) does not increase the ϵ-capacity, namely,

$$\lim_{\ell\to\infty}\frac{1}{\ell}\log M_t^*(\ell,\epsilon) = [\![C_\epsilon]\!]\,, \tag{20}$$

where $[\![C_\epsilon]\!]$ is defined in (12) and given by Theorem 1.

However, a much more important implication of Theorem 2 is the following. If we denote by $M^*(n,\epsilon)$ the fundamental limit of coding with fixed blocklength and no feedback (which is equal to the maximal cardinality of the code with blocklength n and probability of error ϵ), then for several channels, including DMCs, the additive white Gaussian noise (AWGN) channel and some channels with memory the behavior of this function at fixed ϵ and moderate n is tightly characterized by the expansion [12], [18]

$$\log M^*(n,\epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n)\,, \tag{21}$$

where C is the channel capacity, V is the channel dispersion and Q^{-1} is the inverse of the standard Q-function:

$$Q(x) = \int_x^\infty \frac{e^{-y^2}}{\sqrt{2\pi}}dy\,. \tag{22}$$

Thus in the absence of feedback the backoff from ϵ-capacity (equal to capacity for DMCs) is governed by the $\frac{1}{\sqrt{n}}$ term (21). The key advantage of variable-length coding with feedback lies in completely eliminating that penalty, thereby opening the possibility of attaining the capacity at a much smaller (average) blocklength.

Furthermore, the achievability (lower) bound in (18) is obtained via stop-feedback codes that use feedback only to let the encoder know that the decoder has made its final decision; namely, the encoder maps f_n satisfy (10). As (18) demonstrates, such a sparing use of feedback does not lead to any significant loss in rate even non-asymptotically. Naturally, such a strategy is eminently practical in many applications, unlike those strategies that require full, noiseless, instantaneous feedback. In the particular case of the BSC, a lower bound (18) with a weaker $\log \ell$ term and with $\frac{\ell C}{1-\epsilon}$ replaced by ℓC has been claimed in [8].

B. Achievability bound

The proof of Theorem 2 relies on a general achievability bound:

Theorem 3: Fix a real number $\gamma > 0$, a channel $\{P_{Y_i|X_1^i Y_1^{i-1}}\}_{i=1}^{\infty}$ and an arbitrary process $X = (X_1, X_2, \ldots, X_n, \ldots)$ taking values in $\mathcal{A}$. Define a probability space with finite-dimensional distributions given by

$$P_{X^n Y^n \bar{X}^n}(a^n, b^n, c^n) = P_{X^n}(a^n) P_{\bar{X}^n}(c^n) \prod_{j=1}^{n} P_{Y_j|X_1^j Y_1^{j-1}}(b_j|a^j, b^{j-1}) \,, \tag{23}$$

i.e. X and $\bar{X}$ are independent copies of the same process and Y is the output of the channel when X is its input. For the joint distribution (23) define a sequence of information density functions $\mathcal{A}^n \times \mathcal{B}^n \to \bar{\mathbb{R}}$

$$\imath(a^n; b^n) = \log \frac{dP_{Y^n|X^n}(b^n|a^n)}{dP_{Y^n}(b^n)} \,, \tag{24}$$

and a pair of hitting times:

$$\tau = \inf\{n \geq 0 : \imath(X^n; Y^n) \geq \gamma\} \,, \tag{25}$$

$$\bar{\tau} = \inf\{n \geq 0 : \imath(\bar{X}^n; Y^n) \geq \gamma\} \,. \tag{26}$$

Then for any M there exists an (ℓ, M, ϵ) VLF code with

$$\ell \leq \mathbb{E}[\tau] \tag{27}$$

$$\epsilon \leq (M-1)\mathbb{P}[\bar{\tau} \leq \tau] \,. \tag{28}$$

Furthermore, for any M there exists a deterministic (ℓ', M, ϵ) VLF code with ϵ satisfying (28) and

$$\ell' \leq \operatorname{esssup} \mathbb{E}\left[\tau|X\right]. \tag{29}$$

Remarks:

1) It is instructive to think of X, Y and $\bar{X}$ as the sent codeword, the output of the channel in response to X and a codeword distributed as X but independent of (X, Y).
2) Worsening the bound to (29) is advantageous, since for symmetric channels we have $\mathbb{E}\left[\tau|X\right] = \mathbb{E}\left[\tau\right]$ and thus the second part of Theorem 3 guarantees the existence of a deterministic code without any sacrifice in performance.
3) Theorem 3 is a natural extension of the DT bound [12, Theorem 17], since (28) corresponds to the second term in [12, (70)], whereas the first term in [12, (70)] is missing because the information density corresponding to the true message eventually crosses any level γ with probability one.
4) Interestingly, pairing a fixed stopping rule with a random-coding argument has been already discovered from a different perspective: in the context of universal variable-length codes [6]–[10], stopping rules based on a sequentially computed EMI were shown to be optimal in several different asymptotic senses. Although invaluable for universal coding, EMI-based decoders are hard to evaluate non-asymptotically and their analysis relies on inherently asymptotic methods, such as type-counting, cf. [10].

Proof: To define a code we need to specify (U, f_n, g_n, τ). First we define a random variable U as follows:

$$\mathcal{U} \stackrel{\triangle}{=} \underbrace{\mathcal{A}^\infty \times \cdots \times \mathcal{A}^\infty}_{M \text{ times}} \tag{30}$$

$$P_U \stackrel{\triangle}{=} \underbrace{P_{X^\infty} \times \cdots \times P_{X^\infty}}_{M \text{ times}}, \tag{31}$$

where P_{X^∞} is the distribution of the process X. Note that even for $|\mathcal{A}| = 2$, $\mathcal{U}$ will have the cardinality of the real line $\mathbb{R}$. However, in view of Theorem 19, $|\mathcal{U}|$ can always be reduced to 3.

The realization of U defines M infinite dimensional vectors $\mathbf{C}_j \in \mathcal{A}^\infty, j = 1, \ldots, M$. Our encoder and decoder will depend on U implicitly through $\{\mathbf{C}_j\}$. The coding scheme consists of a sequence of encoders f_n that map a message j to an infinite sequence of inputs $\mathbf{C}_j \in \mathcal{A}^\infty$

without any regard to feedback:

$$f_n(w) = (\mathbf{C}_w)_n \,, \tag{32}$$

where $(\mathbf{C}_j)_n$ is the n-th coordinate of the vector $\mathbf{C}_j$. Obviously, such encoder satisfies (10).

At time instant n the decoder computes M information densities:

$$S_{j,n} \stackrel{\triangle}{=} \imath(\mathbf{C}_j(n); Y^n) \,, \qquad j = 1, \ldots, M, \tag{33}$$

where $\mathbf{C}_j(n)$ is the restriction of $\mathbf{C}_j$ to the first n symbols. The decoder also defines M stopping times:

$$\tau_j \stackrel{\triangle}{=} \inf\{n \geq 0 : S_{j,n} \geq \gamma)\} \,. \tag{34}$$

The final decision is made by the decoder at the stopping time τ^*:

$$\tau^* \stackrel{\triangle}{=} \min_{j=1,\ldots M} \tau_j \,. \tag{35}$$

This means that τ^* is the moment of the first γ-upcrossing among all S_j. The output of the encoder is

$$g(Y^{\tau^*}) = \max\{j : \tau_j = \tau^*\} \,. \tag{36}$$

We are left with the problem of choosing $\mathbf{C}_j, j = 1, \ldots M$.

This will be done by generating $\mathbf{C}_j$ randomly, independently of each other and distributed according to P_{X^∞} on $\mathcal{A}^\infty$.

We give an interpretation for our decoding scheme in the special case of a memoryless channel with $P_{X^\infty} = P_X^\infty$, i.e. X_k are independent and identically distributed with a single-letter distribution P_X. In this case, the decoder observes M random walks S_j one of which has a positive drift $I(X;Y)$ (the true message) and $(M-1)$ have negative drifts $-D(P_X P_Y || P_{XY})$, a quantity known as lautum information $L(X;Y)$, see [22]. The goal of the decoder, of course, is to detect the one with positive drift.

The average length of transmission satisfies:

$$\begin{aligned} \mathbb{E}\left[\tau^*\right] &\leq \frac{1}{M}\sum_{j=1}^{M} \mathbb{E}\left[\tau_j | W = j\right] && (37)\\ &= \mathbb{E}\left[\tau_1 | W = 1\right] && (38)\\ &= \mathbb{E}\left[\tau\right], && (39) \end{aligned}$$

where (38) is by symmetry and (39) follows by the definition of τ in (25). Analogously, the average probability of error satisfies

$$\begin{aligned}
\mathbb{P}[g(Y^{\tau^*}) \neq W] &\leq \mathbb{P}[g(Y^{\tau^*}) \neq 1 | W=1] && (40)\\
&\leq \mathbb{P}[\tau_1 \geq \tau^* | W=1] && (41)\\
&\leq \mathbb{P}\left[\bigcup_{j=2}^{M} \{\tau_j \leq \tau_1\} \middle| W=1\right] && (42)\\
&\leq (M-1)\mathbb{P}[\tau_2 \leq \tau_1 | W=1]\,, && (43)
\end{aligned}$$

where (40) is by (36), (42) is by the definition (35), and (43) is by a union bound and symmetry. Finally, notice that conditioned on $W=1$ the joint distribution of $(S_{1,n}, S_{2,n}, \tau_1, \tau_2)$ is exactly the same as that of $(\imath(X^n;Y^n), \imath(\bar{X}^n;Y^n), \tau, \bar{\tau})$ defined in the formulation of the theorem and (25), thus we have proved (27) and (28).

To prove (29) simply notice that similarly to (39) we have almost surely:

$$\mathbb{E}\left[\tau^* | U\right] \leq \operatorname{esssup} \mathbb{E}\left[\tau | X\right], \tag{44}$$

and thus the bound (29) is automatically satisfied for every realization U. On the other hand, because of (43) there must exist a realization u_0 of U such that

$$\mathbb{P}[g(Y^{\tau^*}) \neq W | U=u_0] \leq (M-1)\mathbb{P}[\bar{\tau} \leq \tau]\,, \tag{45}$$

which therefore defines a deterministic code with the sought-after performance (28) and (29). ■

C. Converse bounds

The converse parts of Theorem 2 follow from the following result:

Theorem 4: Consider an arbitrary DMC with capacity C. Then any (ℓ, M, ϵ) VLF code with $0 \leq \epsilon \leq 1 - \frac{1}{M}$ satisfies

$$\log M \leq \frac{C\ell + h(\epsilon)}{1-\epsilon}\,, \tag{46}$$

whereas each (ℓ, M, ϵ) VLFT code with $0 \leq \epsilon \leq 1 - \frac{1}{M}$ satisfies

$$\begin{aligned}
\log M &\leq \frac{C\ell + h(\epsilon) + (\ell+1)h\left(\frac{1}{\ell+1}\right)}{1-\epsilon} && (47)\\
&\leq \frac{C\ell + \log(\ell+1) + h(\epsilon) + \log e}{1-\epsilon}\,, && (48)
\end{aligned}$$

where $h(x) = -x\log x - (1-x)\log(1-x)$ is the binary entropy function.

Proof: The inequality (46) is contained essentially in Lemmas 1 and 2 of [3]. Thus we focus on (47) only briefly mentioning how to obtain (46). First we give an informal argument. According to the Fano inequality

$$\begin{aligned}(1-\epsilon)\log M &\leq I(W;Y^\tau,\tau)+h(\epsilon) && (49)\\ &= I(W;Y^\tau)+I(W;\tau|Y^\tau)+h(\epsilon) && (50)\\ &\leq I(W;Y^\tau)+H(\tau)+h(\epsilon) && (51)\\ &\leq I(W;Y^\tau)+(\ell+1)h\left(\frac{1}{\ell+1}\right)+h(\epsilon) && (52)\\ &\leq C\ell+(\ell+1)h\left(\frac{1}{\ell+1}\right)+h(\epsilon)\,, && (53)\end{aligned}$$

where in (52) we have upper-bounded $H(\tau)$ by solving a simple optimization problem[3] for an integer-valued non-negative random variable τ:

$$\max_{\tau:\mathbb{E}[\tau]\leq\ell} H(\tau)=(\ell+1)h\left(\frac{1}{\ell+1}\right)\,, \tag{54}$$

and in (53) we used the result of Burnashev [3]:

$$I(W;Y^\tau)\leq C\,\mathbb{E}[\tau]\leq C\ell\,. \tag{55}$$

Clearly (53) is equivalent to (47). The case of VLF codes is even simpler since τ is a function of Y^τ and thus $I(W;Y^\tau,\tau)=I(W;Y^\tau)$.

Unfortunately, the random variables (Y^τ,τ) and Y^τ are not well-defined and thus a different proof is required. Nevertheless, the main idea still pivots on the fact that because of the restriction on expectation, τ cannot convey more than $O(\log\ell)$ bits of information about the message.

Initially, we will assume that the code is deterministic and $|U|=1$. Consider a triplet (f_n,g_n,τ) defining a given code. For a VLFT code, τ is a stopping moment of the filtration $\sigma\{W,Y^k\}_{k=0}^\infty$. To get rid of dependence of τ on W we introduce an extended channel $(\hat{\mathcal{A}},\hat{\mathcal{B}},P_{\hat{Y}|\hat{X}})$ as follows:

$$\begin{aligned}\hat{\mathcal{A}} &= \mathcal{A}\cup\{T\}\,, && (56)\\ \hat{\mathcal{B}} &= \mathcal{B}\cup\{T\}\,, && (57)\\ P_{\hat{Y}|\hat{X}}(\hat{y}|\hat{x}) &= \begin{cases}P_{Y|X}(\hat{y}|\hat{x})\,, & \hat{x}\neq T\,,\\ 1\{\hat{y}=T\}\,, & \hat{x}=T\,.\end{cases} && (58)\end{aligned}$$

[3]The solution is given by a geometric distribution.

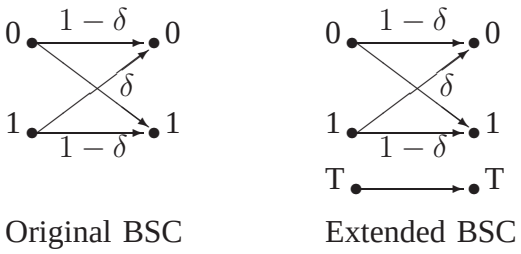

Fig. 1. Illustration of the channel extension in the proof of Theorem 4.

In other words, the channel $P_{\hat{Y}|\hat{X}}$ has an additional input T conveyed noiselessly to the output. If $P_{Y|X}$ is a BSC with crossover probability δ then the extended channel has transition diagram as represented on Fig. 1.[4] We also assume that the original and extended channels are defined on the same probability space where they are coupled in such a way that whenever $\hat{X} = X$ we have $\hat{Y} = Y$.

Next, we convert the given code (f_n, g_n, τ) to the code $(\hat{f}_n, \hat{g}_n, \hat{\tau})$ for the extended channel as follows:

$$\hat{f}_n(W, \hat{Y}^{n-1}) = \begin{cases} f_n(W, \hat{Y}^{n-1}), & \tau \geq n, \\ T, & \tau < n, \end{cases} \tag{59}$$

$$\hat{\tau} = \tau + 1 = \inf\{n : \hat{Y}_n = T\}, \tag{60}$$

$$\hat{g}_n(\hat{Y}^n) = \begin{cases} g_n(\hat{Y}^n), & \hat{\tau} > n, \\ g_n(\hat{Y}^{\hat{\tau}-1}) & \hat{\tau} \leq n, \end{cases}. \tag{61}$$

Note that by definition $\tau \geq n$ can be decided by knowing W and Y^{n-1} only and hence $\hat{f}_n$ is indeed a function of $(W, \hat{Y}^{n-1})$; also notice that $\hat{Y}^{n-1} \in \mathcal{A}^{n-1}$ whenever $\tau \geq n$, and therefore the expression $f_n(W, \hat{Y}^{n-1})$ is meaningful.

Since $\hat{\tau}$ is a stopping time of the filtration

$$\mathscr{F}_n \stackrel{\triangle}{=} \sigma\{\hat{Y}^j\}_{j=1}^n \tag{62}$$

the triplet $(\hat{f}_n, \hat{g}_n, \hat{\tau})$ forms an $(\ell + 1, M, \epsilon)$ VLF code for the extended channel (58). This code satisfies an additional constraint: input symbol T is used only once and it terminates the transmission. Now we prove that any such code must satisfy a certain upper bound on its

[4]The extended BSC was the first DMC to be analyzed in information theory [13].

cardinality M. To do so, consider the space $\{1,\dots,M\}\times\hat{\mathcal{A}}^\infty$ and two measures on it: $P_{W\hat{Y}^\infty}$ and $P_W\times P_{\hat{Y}^\infty}$, where $P_{W\hat{Y}^\infty}$ is the joint distribution of random variables W and $\hat{Y}^\infty$ induced by the code $(\hat{f}_n,\hat{g}_n,\hat{\tau})$. Consider a measurable function

$$\phi:\{1,\dots,M\}\times\hat{\mathcal{A}}^\infty\to\{0,1\} \tag{63}$$

defined as

$$\phi=1\{\hat{g}_{\hat{\tau}}(Y^{\hat{\tau}})=W\}\,. \tag{64}$$

Notice that under measure $P_{W\hat{Y}^\infty}$ we have:

$$P_{W\hat{Y}^\infty}[\phi=1]\geq 1-\epsilon\,, \tag{65}$$

due to the requirement (6). On the other hand, since under $P_W\times P_{\hat{Y}^\infty}$ $\hat{g}_{\hat{\tau}}$ is independent of W, we have

$$(P_W\times P_{\hat{Y}^\infty})[\phi=1]=\frac{1}{M}\,. \tag{66}$$

By assumption $1-\epsilon\geq\frac{1}{M}$ and therefore by the data-processing inequality we must have

$$D(P_{W\hat{Y}^\infty}||P_WP_{\hat{Y}^\infty})\geq d(1-\epsilon||\tfrac{1}{M})\,, \tag{67}$$

where $d(x||y)=x\log\frac{x}{y}+(1-x)\log\frac{1-x}{1-y}$ is the binary relative entropy. After straightforward manipulations in (67) we obtain

$$(1-\epsilon)\log M\leq I(W;\hat{Y}^\infty)+h(\epsilon)\,. \tag{68}$$

Although, (68) is just the Fano inequality, inclusion of the complete derivation illustrates the similarity with the meta-converse approach in Theorem 26 and Section III.G in [12]. Another important observation is that for small ℓ, the bound can be tightened by replacing the step of data-processing (67) with an exact non-asymptotic solution of the Wald's sequential hypothesis testing problem.

We proceed to upper bound $I(W;\hat{Y}^\infty)$.[5] To do so we define a sequence of random variables:

$$Z_k=\log\frac{P_{\hat{Y}_k|W\hat{Y}^{k-1}}(\hat{Y}_k|W,\hat{Y}^{k-1})}{P_{\hat{Y}_k|\hat{Y}^{k-1}}(\hat{Y}_k|\hat{Y}^{k-1})}\,, \tag{69}$$

[5]Notice that $\hat{Y}^\infty$ formalizes the idea of viewing (Y^τ,τ) as a random variable.

which are relevant to $I(W;\hat{Y}^\infty)$ because by simple telescoping we have

$$\begin{aligned} I(W;\hat{Y}^\infty) &= I(W;\hat{Y}_1) + I(W;\hat{Y}_2^\infty|\hat{Y}_1) && (70)\\ &= \sum_{k=1}^{\infty} I(W;\hat{Y}_{k+1}|Y^k) && (71)\\ &= \sum_{k=1}^{\infty} \mathbb{E}\left[Z_k\right]. && (72)\end{aligned}$$

For Z_k we have the following property:

$$\mathbb{E}\left[Z_k|\mathscr{F}_{k-1}\right] = I_{\mathscr{F}_{k-1}}(W;\hat{Y}_k)\,, \tag{73}$$

where $I_{\mathscr{F}}(\cdot;\cdot)$ denotes mutual information, conditioned on $\mathscr{F}$. Specifically, for discrete random variables A, B and C we define the following $\mathscr{F}$-measurable random variable:

$$I_{\mathscr{F}}(A;B|C) = \sum_{a,b,c} \mathbb{P}[A=a,B=b,C=c\,|\,\mathscr{F}] \log \frac{\mathbb{P}[A=a,B=b,C=c\,|\,\mathscr{F}]\,\mathbb{P}[C=c|\mathscr{F}]}{\mathbb{P}[A=a,C=c\,|\,\mathscr{F}]\,\mathbb{P}[B=b,C=c\,|\,\mathscr{F}]}\,, \tag{74}$$

where the summation is over the alphabets of A, B and C. We also define

$$H_{\mathscr{F}}(A) \stackrel{\triangle}{=} I_{\mathscr{F}}(A;A)\,, \tag{75}$$

and other information measures similarly.

We define yet another process adapted to filtration $\mathscr{F}_n$, cf. (62),

$$V_n \stackrel{\triangle}{=} 1\{\hat{\tau} \le n\}\,. \tag{76}$$

With this notation we have:

$$\begin{aligned} I_{\mathscr{F}_{k-1}}(W;\hat{Y}_k) &= I_{\mathscr{F}_{k-1}}(W;\hat{Y}_k V_k) && (77)\\ &= I_{\mathscr{F}_{k-1}}(W;V_k) + I_{\mathscr{F}_{k-1}}(W;\hat{Y}_k|V_k) && (78)\\ &\le H_{\mathscr{F}_{k-1}}(V_k) + I_{\mathscr{F}_{k-1}}(W;\hat{Y}_k|V_k) && (79)\\ &\le H_{\mathscr{F}_{k-1}}(V_k) + I_{\mathscr{F}_{k-1}}(\hat{X}_k;\hat{Y}_k|V_k)\,, && (80)\end{aligned}$$

where (77) follows because V_k is a function of $\hat{Y}_k$, (78) is the usual chain rule and (80) is obtained by applying the data-processing lemma to the Markov relation $W - \hat{X}_k - \hat{Y}_k - V_k$, which holds almost surely when conditioned on $\mathscr{F}_{k-1}$. We now upper-bound the second term in (80) as follows

$$I_{\mathscr{F}_{k-1}}(\hat{X}_k;\hat{Y}_k|V_k) \le 0\cdot\mathbb{P}[V_k=1|\mathscr{F}_{k-1}] + \mathbb{P}[V_k=0|\mathscr{F}_{k-1}]C\,, \tag{81}$$

because when $V_k = 1$ we must have $\hat{X}_k = \hat{Y}_k = T$ and the mutual information is zero, while when $V_k = 0$ we are computing the mutual information acquired on the $P_{\hat{Y}|\hat{X}}$ channel over a distribution $P_{\hat{X}_k|V_k \neq 0}$ which has a zero mass on the symbol T, and thus

$$\sup_{P_{\hat{X}}: P_{\hat{X}}(T)=0} I(\hat{X}; \hat{Y}) = C \,. \tag{82}$$

Overall, from (73), (80) and (81) it follows:

$$\mathbb{E}\left[Z_k | \mathscr{F}_{k-1}\right] \leq H_{\mathscr{F}_{k-1}}(V_k) + \mathbb{P}[V_k = 0 | \mathscr{F}_{k-1}] C \,. \tag{83}$$

Finally, we obtain

$$\begin{aligned}
I(W; \hat{Y}^\infty) &= \sum_{k=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[Z_k | \mathscr{F}_{k-1}\right]\right] && (84)\\
&\leq \sum_{k=1}^{\infty} H(V_k | \hat{Y}^{k-1}) + \mathbb{P}[V_k = 0] C && (85)\\
&= \sum_{k=1}^{\infty} H(V_k | \hat{Y}^{k-1}) + C\,\mathbb{E}\left[\tau\right] && (86)\\
&\leq \sum_{k=1}^{\infty} H(V_k | V^{k-1}) + C\,\mathbb{E}\left[\tau\right] && (87)\\
&= H(V_1, V_2, \ldots) + C\,\mathbb{E}\left[\tau\right] && (88)\\
&= H(\hat{\tau}) + C\,\mathbb{E}\left[\tau\right] && (89)\\
&= H(\tau) + C\,\mathbb{E}\left[\tau\right] && (90)
\end{aligned}$$

where (84) follows from (72), (85) results from (83), (86) follows by taking an expectation of the obvious identity

$$\sum_{k=1}^{\infty} 1\{V_k = 0\} = \sum_{k=1}^{\infty} 1\{\hat{\tau} > k\} = \hat{\tau} - 1 \,, \tag{91}$$

and recalling that $\hat{\tau} - 1 = \tau$, (87) follows because V^{k-1} is a function of $\hat{Y}^{k-1}$, (88) is obtained by the entropy chain rule, (90) follows since $(V_1, V_2, \ldots, V_n, \ldots)$ is an invertible function of $\hat{\tau}$, and finally (90) follows since $\hat{\tau} = \tau + 1$.

Together (68), (90) and (54) prove (47) in the case of a deterministic code with $|U| = 1$. For the case of $|U| > 1$ the above argument has shown that we have

$$(1 - \mathbb{P}[W \neq \hat{W}|U]) \log M \leq C\,\mathbb{E}\left[\tau | U\right] + H_{\sigma\{U\}}(\tau) + h(\mathbb{P}[W \neq \hat{W}|U]) \qquad \text{a.s.,} \tag{92}$$

where $\hat{W} = g_\tau(Y^\tau)$ is the output message estimate of the decoder. By taking the expectation of both sides of (92) and applying the Jensen's inequality to the binary entropy terms we obtain

$$(1 - \mathbb{P}[W \neq \hat{W}]) \log M \leq C\, \mathbb{E}\,[\tau] + H(\tau|U) + h(\epsilon)\,, \tag{93}$$

and then (47) follows since by (54) we have

$$H(\tau|U) \leq H(\tau) \leq (\ell+1) h\left(\frac{1}{\ell+1}\right)\,. \tag{94}$$

Notice that in the case of VLF codes, the first term in (86) disappears because V_k is a function of $\hat{Y}^{k-1}$ thus leading to the tighter bound (46). ■

An alternative to the converse in (46) for channels with $C_1 < \infty$ was discovered by Burnashev [3, Theorem 1] in order to show optimality of the exponent (1). A stronger version of that result with a streamlined proof was given in [14]:

Theorem 5 ([14]): Consider a DMC with $0 < C \leq C_1 < \infty$. Then any (ℓ, M, ϵ) VLF code satisfies

$$\ell \geq \sup_{0<\xi\leq\frac{1}{2}} \left[\left(1 - \xi - \frac{\epsilon}{\xi}\right)\frac{\log M}{C} + \frac{1}{C_1}\log\frac{\lambda\xi}{4\epsilon} - \frac{h(\xi)}{C}\right]\,, \tag{95}$$

where

$$C_1 \;=\; \max_{a_1,a_2\in\mathcal{A}} D(P_{Y|X=a_1}||P_{Y|X=a_2}) \tag{96}$$

$$\lambda \;\stackrel{\triangle}{=}\; \min_{x,y} P_{Y|X}(y|x) > 0\,. \tag{97}$$

The proofs of both [3, Theorem 1] and Theorem 5 rely on seminal ideas of [15] and [3], who proposed to split the analysis of a given code in two phases using an auxiliary stopping time $\tau_1 \leq \tau$. Burnashev used τ_1 defined as the first time when the conditional entropy $H(W|Y^n)$ falls below a threshold $A > 0$. Instead, [15] proposed τ_1 to be the first time when $\max_w P_{W|Y^n}(w|Y^n)$ reaches a threshold $1-\xi$. As demonstrated in [14], such a choice results in a much more elegant proof. Note that unlike [14], the original result in [15] was asymptotic, and restricted to the case of the AWGN channel. Moreover the reasoning in [15] contained a flaw, as pointed out by Burnashev [3].

One drawback of the bound (95) is that it is not always stronger than (46). For example, for a capacity-$\frac{1}{2}$ BSC and $\epsilon = 10^{-3}$, (95) is worse than (46) for all delays. To rectify this situation we give a new bound which is provably tighter than both (46) and Theorem 5. The proof, included in the appendix, employs the two-phase approach choosing the same τ_1 as in [14],

[15]. Furthermore, it follows the meta-converse framework of [12, Section III.E] and [16, Section 2.7].

Theorem 6: Consider a DMC with $0 < C \le C_1 < \infty$. Then any (ℓ, M, ϵ) VLF code with $0 < \epsilon \le 1 - \frac{1}{M}$ satisfies

$$\ell \ge \sup_{0<\xi\le 1-\frac{1}{M}} \left[\frac{1}{C} \left(\log M - F_M(\xi) - \min \left\{ F_M(\epsilon), \frac{\epsilon}{\xi} \log M \right\} \right) + \left| \frac{1-\epsilon}{C_1} \log \frac{\lambda_1 \xi}{\epsilon(1-\xi)} - \frac{h(\epsilon)}{C_1} \right|^+ \right], \tag{98}$$

where

$$F_M(x) \triangleq x \log(M-1) + h(x), \quad 0 \le x \le 1 \tag{99}$$

$$\lambda_1 \triangleq \min_{y,x_1,x_2} \frac{P_{Y|X}(y|x_1)}{P_{Y|X}(y|x_2)} \in (0,1). \tag{100}$$

Numerical experimentation suggests that weakening (98) by replacing the minimum by $F_M(\epsilon)$ has negligible effect.

D. Asymptotic expansions

Proof of Theorem 2: The upper bounds in (16) and (17) follow from Theorem 4. For the lower bound (16), suppose that for each ℓ' there exists an $\left(\ell', M, \frac{1}{\ell'}\right)$-VLF code with

$$\log M = C\ell' - \log \ell' - a_0, \tag{101}$$

where a_0 is some constant. To see that (101) implies the lower bound in (16) consider the code which terminates without any channel uses, i.e. $\tau = 0$, with probability $\frac{\ell'\epsilon - 1}{\ell' - 1}$ and uses the $\left(\ell', M, \frac{1}{\ell'}\right)$-VLF code otherwise[6]. Such a code has probability of error ϵ and average length $\ell = \frac{\ell'^2(1-\epsilon)}{\ell'-1}$ and, therefore, using (101) we have

$$\log M^*(\ell, \epsilon) \ge C\ell' - \log \ell' - a_0 \tag{102}$$

$$= \frac{\ell C}{1-\epsilon} - \log \ell + O(1), \tag{103}$$

as required.

[6]Note that due to availability of the stop feedback such a randomization can be realized on the decoder side only, i.e. without requiring any common randomness, U. Thus if $\left(\ell', M, \frac{1}{\ell'}\right)$-VLF code exists with $|U| = 1$ then the overall coding scheme constructed to achieve (16) also has $|U| = 1$.

To prove (101), we apply Theorem 3 with the process $\{X_n\}_{n=1}^{\infty}$ chosen to be independent and identically distributed (i.i.d.) with a marginal distribution P_X – a capacity achieving distribution. To analyze (28) it is convenient to define a pair of random walks

$$S_n \stackrel{\triangle}{=} \imath(X^n;Y^n)\,, \tag{104}$$

$$\bar{S}_n \stackrel{\triangle}{=} \imath(\bar{X}^n;Y^n)\,. \tag{105}$$

First notice that since the sequence $S_n - nI(X;Y) = S_n - nC$ is a martingale we obtain from Doob's optional stopping theorem [17, Theorem 10.10]

$$C\,\mathbb{E}\,[\tau] = \mathbb{E}\,[S_\tau] \tag{106}$$

$$\leq \gamma + a_0\,, \tag{107}$$

where a_0 is an upper-bound on S_1. The equality

$$D(P||Q)\mathbb{E}\,[\tau] = \mathbb{E}\left[\log\frac{dP}{dQ}\bigg|_{\mathscr{F}_\tau}\right] \tag{108}$$

is traditionally called Wald's identity in the sequential hypothesis testing literature. In particular, we obtain from (107)

$$\mathbb{P}[\tau<\infty] = 1 \tag{109}$$

Next notice that for any (measurable) function f we have

$$\mathbb{E}\,[f(\bar{X}^n,Y^n)] = \mathbb{E}\,[f(X^n,Y^n)\exp\{-S_n\}]\,, \tag{110}$$

because $S_n = \log\frac{dP_{X^nY^n}}{dP_{\bar{X}^nY^n}}$. Therefore, we have

$$\mathbb{P}[\bar{\tau}\leq\tau] \leq \mathbb{P}[\bar{\tau}<\infty] \tag{111}$$

$$= \lim_{n\to\infty}\mathbb{P}[\bar{\tau}<n] \tag{112}$$

$$= \lim_{n\to\infty}\mathbb{E}\,[\exp\{-S_n\}1\{\tau<n\}] \tag{113}$$

$$= \lim_{n\to\infty}\mathbb{E}\,[\exp\{-S_n\}1\{\tau_n<n\}] \tag{114}$$

$$= \lim_{n\to\infty}\mathbb{E}\,[\exp\{-S_{\tau_n}\}1\{\tau_n<n\}] \tag{115}$$

$$= \mathbb{E}\,[\lim_{n\to\infty}(\exp\{-S_{\tau_n}\}1\{\tau_n<n\})] \tag{116}$$

$$= \mathbb{E}\,[\exp\{-S_\tau\}1\{\tau<\infty\}] \tag{117}$$

$$\leq \exp\{-\gamma\}\,, \tag{118}$$

where (111) is from (109), (112) is by monotonicity, (113) is from (110), in (114) we have defined

$$\tau_n \stackrel{\triangle}{=} \min\{\tau, n\}\,, \tag{119}$$

from which (114) follows, (115) is by the optional stopping theorem [17, Theorem 10.10] applied to the martingale $\exp\{-S_n\}$ and stopping time τ_n, and finally (116) and (118) both follow from

$$\exp\{-S_{\tau_n}\}1\{\tau_n < n\} = \exp\{-S_\tau\}1\{\tau_n < n\} \le \exp\{-\gamma\}\,, \tag{120}$$

which in turn follows from the definition of τ in (25).

The existence of an $\left(\ell', M, \frac{1}{\ell'}\right)$-VLF code with M satisfying (101) now follows by taking $\gamma = C\ell' - a_0$ and using (107) and (118) in (27) and (28), respectively. ■

We note in passing that while the codes with encoders utilizing full noiseless feedback can achieve the Burnashev exponent (1), it was noted in [8], [10] that the lower error exponent

$$E_1(R) = C - R \tag{121}$$

is achievable at all rates $R < C$ with stop-feedback codes (10). Indeed, this property easily follows from (118) and (107).

A numerical comparison of the upper and lower bounds for the BSC with crossover probability $\delta = 0.11$ and $\epsilon = 10^{-3}$ is given in Fig. 2, where the upper bound is (98) and the lower bound is Theorem 3 evaluated for various M and the lowest possible γ for which the right-hand side of (28) is still smaller than 10^{-3}. Note that for $BSC(\delta)$ the $\imath(X^n;Y^n)$ becomes a random walk taking steps $\log 2\delta$ and $\log(2-2\delta)$ with probabilities δ and $1-\delta$, i.e.,

$$\imath(X^n;Y^n) = n\log(2-2\delta) + \log\frac{\delta}{1-\delta}\sum_{k=1}^{n} Z_k\,, \tag{122}$$

where Z_k are independent Bernoulli $\mathbb{P}[Z_k = 1] = 1 - \mathbb{P}[Z_k = 0] = \delta$. The evaluation of (28) is simplified by using (110) to get rid of the process $\imath(\bar{X}^n;Y^n)$, which in this case is independent of (X^n, Y^n):

$$\epsilon \le (M-1)\mathbb{E}\left[f(\tau)\right], \tag{123}$$

where

$$f(n) \stackrel{\triangle}{=} \mathbb{E}\left[1\{\tau \le n\}\exp\{-\imath(X^\tau;Y^\tau)\}\right]. \tag{124}$$

The dashed line in Fig. 2 is the approximate fundamental limit for fixed blocklength codes without feedback given by the equation (21) with $O(\log n)$ substituted by $\frac{1}{2}\log n$; see [12, Theorem 53].

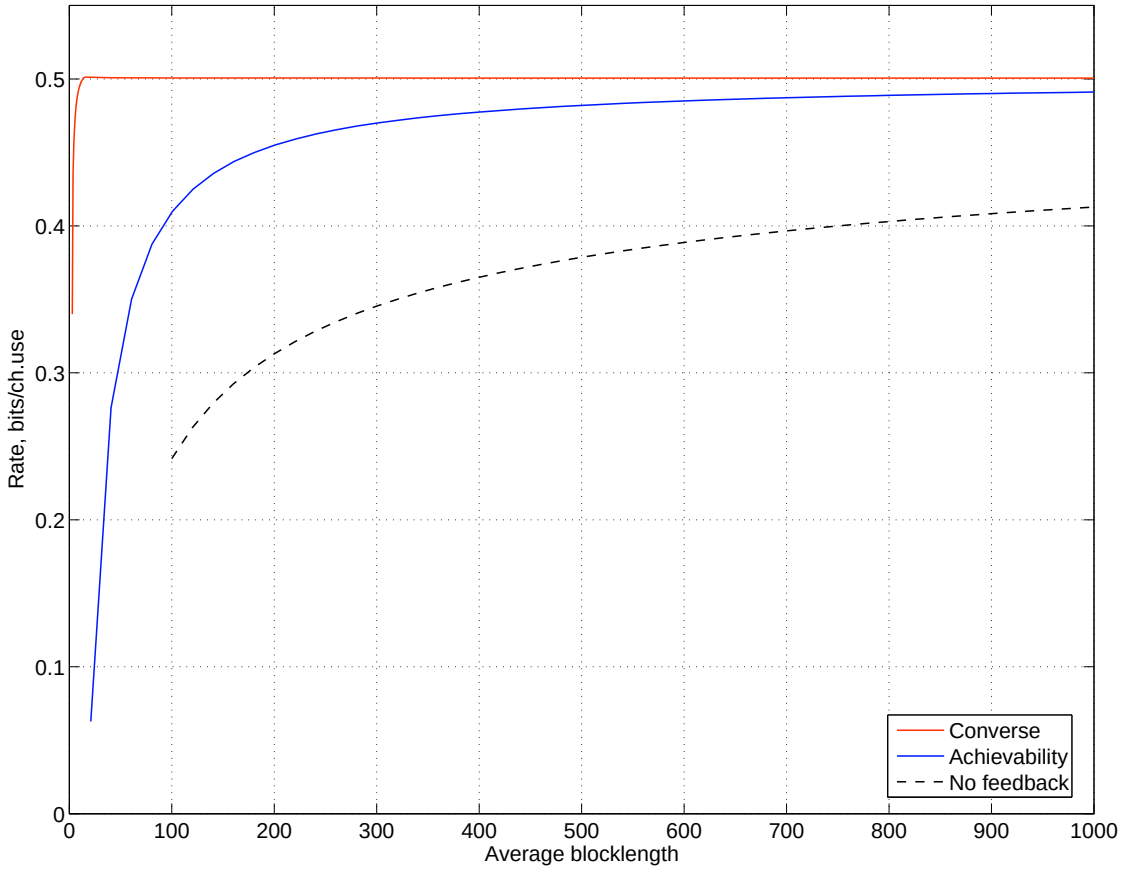

Fig. 2. Comparison of upper and lower bounds on the maximal achievable rate of variable-length feedback coding for the BSC(0.11); probability of error $\epsilon = 10^{-3}$.

Theorem 7: For a $BEC(\delta)$ and $\epsilon \in [0, 1)$ we have

$$\log_2 M_f^*(\ell, \epsilon) = \frac{\ell C}{1-\epsilon} + O(1)\,, \tag{125}$$

where $C = 1 - \delta$ bit. More precisely,

$$\left\lfloor \frac{\ell C}{1-\epsilon} \right\rfloor \le \log_2 M_f^*(\ell, \epsilon) \le \frac{\ell C}{1-\epsilon} + \frac{h(\epsilon)}{1-\epsilon}\,. \tag{126}$$

Proof: The upper bound in Theorem 2 holds even for $\epsilon = 0$, so we need only to prove a lower bound. First, we assume $\epsilon = 0$ and take arbitrary k. Consider the strategy that simply retransmits each of k bits until it gets through the channel unerased. More formally, we define a stopping time as

$$\tau_0 = \inf\{n \ge 1 : \text{ there are } k \text{ unerased symbols in } Y_1, \dots Y_n\}\,. \tag{127}$$

It is easy to see that

$$\mathbb{E}\left[\tau_0\right] = \frac{k}{1-\delta}\,. \tag{128}$$

Hence for any ℓ we have shown

$$\log_2 M_f^*(\ell, 0) \geq \lfloor \ell C \rfloor\,. \tag{129}$$

For $\epsilon > 0$ we make use of the randomization to construct a transmission scheme that stops at time 0 with probability ϵ and otherwise proceeds as above. We define a stopping time

$$\tau_\epsilon = \tau_0 1\{U \geq \epsilon\}\,, \tag{130}$$

where U is uniform on $[0,1]$ and measurable with respect to $\mathscr{G}_0$. It is clear that using such a strategy we obtain a probability of error upper-bounded by ϵ and

$$\mathbb{E}\left[\tau_\epsilon\right] = \frac{k}{1-\delta}(1-\epsilon)\,. \tag{131}$$

Hence we are able to achieve

$$\log_2 M_f^*(\ell, \epsilon) \geq \left\lfloor \frac{\ell C}{1-\epsilon} \right\rfloor\,. \tag{132}$$

■

The result of Theorem 7 suggests that to improve the expansion (16) to the order $O(1)$, it is likely that we need to go beyond encoders satisfying (10). In the problem of achieving the optimal error exponent, similar reasons necessitate going beyond stop feedback and lead to introducing a second communication phase as in [3] and [5].

IV. Zero-error communication

The general achievability bound, Theorem 3, applies only to $\epsilon > 0$. What can be said about $\epsilon = 0$?

A. No termination symbol (VLF codes)

Burnashev [3] showed that if $C_1 = \infty$, then as $\ell \to \infty$ we have for some $a > 0$

$$\log M_f^*(\ell, 0) \geq \ell C - a\sqrt{\ell \log \ell} + O(\log \ell)\,. \tag{133}$$

For this reason, for such channels zero-error VLF capacity is equal to the conventional capacity. However, the bound $\sqrt{\ell \log \ell}$ on the penalty term is rather loose, as the following result demonstrates.

Theorem 8: For a $BEC(\delta)$ with capacity $C = 1 - \delta$ bit we have

$$\log_2 M_f^*(\ell, 0) = \ell C + O(1)\,. \tag{134}$$

Proof: Theorem 7 applied with $\epsilon = 0$. ■

Regarding any channel with $C_1 < \infty$ (e.g. the BSC), the following negative result holds:

Theorem 9: For any DMC with $C_1 < \infty$ we have

$$\log M_f^*(\ell, 0) = 0 \tag{135}$$

for all $\ell \geq 0$.

Proof: We show that when $C_1 < \infty$ no $(\ell, 2, 0)$ VLF code exists. Indeed, assume that (U, f_n, g_n, τ) is such a code. For zero-error codes, randomization cannot help[7] and hence, without loss of generality we assume $|\mathcal{U}| = 1$. The result can now be derived from [3, Theorem 1], from (95) (both applicable to $|\mathcal{U}| = 1$) or from (98) by noticing that any $(\ell, M, 0)$ VLF code is also an (ℓ, M, ϵ) code for any $\epsilon > 0$ and taking $\epsilon \to 0$. However, it is instructive to give an independent direct proof, which generalizes to infinite alphabets and channels with memory.

Conditioning on $W = 1$ and $W = 2$ gives two measures P_1 and P_2 on B, which are mutually singular when considered on the σ-algebra $\mathscr{G}_\tau$, where $\mathscr{G}_n = \sigma\{Y_1, \ldots, Y_n\}$ is a filtration on B, with respect to which τ is a stopping time. Define a process, adapted to the same filtration:

$$R_n = \log \left.\frac{dP_1}{dP_2}\right|_{\mathscr{G}_n}\,, \tag{136}$$

where $\left.\frac{dP_1}{dP_2}\right|_{\mathscr{G}_n}$ denotes the Radon-Nikodym derivative between P_1 and P_2 considered as measures on the space B with σ-algebra $\mathscr{G}_n$. Then, by memorylessness we have

$$R_n - R_{n-1} = \log \frac{P_{Y|X}(Y_n|f_n(1, Y^{n-1}))}{P_{Y|X}(Y_n|f_n(2, Y^{n-1}))}\,. \tag{137}$$

From (137) and $C_1 < \infty$ it follows that there exists a constant $a_1 > 0$ such that

$$R_n - R_{n-1} \geq -a_1\,, \tag{138}$$

and, consequently,

$$R_n \geq -na_1\,. \tag{139}$$

[7]Indeed, for each u_0 we must have $\mathbb{P}[W \neq \hat{W}|U = u_0] = 0$ and thus we can take the value u_0 which minimizes $\mathbb{E}\,[\tau|U = u_0]$.

On the other hand, taking the conditional expectation of (137) with respect to P_1 we obtain from the definition of C_1 in (96):

$$\mathbb{E}\left[R_n|\mathscr{G}_{n-1}\right] \le R_{n-1} + C_1 < \infty\,, \tag{140}$$

where here and in the remainder of this proof the expectation $\mathbb{E}$ is taken with respect to measure P_1. Thus (140) implies that under P_1 the process $R_n - nC_1$ is a supermartingale. For any integer $k \ge 0$ the random variable $\min\{\tau, k\}$ is a bounded stopping time. Therefore, by Doob's stopping time theorem [17, Theorem 10.10] we have

$$\mathbb{E}\left[R_{\min\{\tau,k\}}\right] \le C_1\mathbb{E}\left[\min\{\tau,k\}\right] \le C_1\mathbb{E}\left[\tau\right] < \infty\,. \tag{141}$$

At the same time, from (139) we have

$$R_{\min\{\tau,k\}} \ge -a_1 \min\{\tau,k\} \ge -a_1\tau\,, \tag{142}$$

and since $\mathbb{E}\left[\tau\right] < \infty$ we can apply Fatou's lemma to (141) to obtain

$$\mathbb{E}\left[R_\tau\right] = \mathbb{E}\left[\liminf_{k\to\infty} R_{\min\{\tau,k\}}\right] \le C_1\mathbb{E}\left[\tau\right] < \infty\,. \tag{143}$$

On the other hand,

$$D_{\mathscr{G}_\tau}(P_1||P_2) = \mathbb{E}\left[R_\tau\right] < \infty\,, \tag{144}$$

thus implying that P_1 and P_2 cannot be mutually singular on $\mathscr{G}_\tau$ – a contradiction. ■

B. Communication with a termination symbol (VLFT codes)

The shortcoming of VLF coding found in Theorem 9 is overcome in the paradigm of VLFT coding. Our main tool is the following achievability bound.

Theorem 10: Fix an arbitrary channel $\{P_{Y_i|X_1^iY_1^{i-1}}\}_{i=1}^\infty$ and a process $X = (X_1, X_2, \ldots, X_n, \ldots)$ with values in $\mathcal{A}$. Then for every positive integer M there exists an $(\ell, M, 0)$ VLFT code with

$$\ell \le \sum_{n=0}^{\infty} \mathbb{E}\left[\min\left\{1, (M-1)\mathbb{P}[\imath(X^n;Y^n) \le \imath(\bar{X}^n;Y^n)|X^nY^n]\right\}\right], \tag{145}$$

where X^n, $\bar{X}^n$, Y^n and $\imath(\cdot;\cdot)$ are defined in (23) and (24). Moreover, this is an FV code which is deterministic and uses feedback only to compute the stopping time, i.e. (10) holds.

Proof: To construct a deterministic code we need to define a triplet (f_n, g_n, τ). Consider a collection of M infinite $\mathcal{A}$-strings $\{\mathbf{C}_1, \ldots, \mathbf{C}_M\}$. The sequence of the encoder functions is defined as

$$f_n(w) = (\mathbf{C}_w)_n \,, \tag{146}$$

where $(\mathbf{C}_j)_n$ is the n-th coordinate of the vector $\mathbf{C}_j$. Recall that in the paradigm of VLFT codes it is allowable for the stopping rule τ to depend on the true message W, so we may define

$$\tau = \inf\{n \geq 0 : \imath(\mathbf{C}_W(n); Y^n) > \max_{v \neq W} \imath(\mathbf{C}_v(n); Y^n)\} \,, \tag{147}$$

where as before $\mathbf{C}_j(n) \in \mathcal{A}^n$ is a restriction of $\mathbf{C}_j$ to the first n coordinates. Definition (147) means that if the true message is j then the transmitter stops at the first time instant n when $\imath(\mathbf{C}_j(n); Y^n)$ is strictly larger than any other $\imath(\mathbf{C}_v(n); Y^n), v \neq j)$. Finally, the sequence of decoder functions is defined as

$$g_n(y^n) = \begin{cases} k, & \text{if } \forall j \neq k : \imath(\mathbf{C}_k(n); y^n) > \imath(\mathbf{C}_j(n); y^n) \\ 1, & \text{otherwise} \,. \end{cases} \tag{148}$$

Upon receiving a stop signal, the decoder outputs the index of the unique message corresponding to the maximal information density, thus we have

$$g_\tau(Y^\tau) = W \,, \tag{149}$$

and the constructed code is indeed a zero-error VLFT code for any selection of M strings $\mathbf{C}_j, j = 1, \ldots, M$. We need to only provide an estimate of the expected length of communication $\mathbb{E}[\tau]$.

The result is proved by applying a random coding argument with each $\mathbf{C}_j$ generated independently with probability distribution P_{X^∞}, corresponding to the fixed input process X. Averaging over all realizations of $\{\mathbf{C}_j, j = 1, \ldots, M\}$ we obtain the following estimate:

$$\begin{aligned} \mathbb{P}[\tau > n] &= \mathbb{P}[\tau > n | W = 1] && (150) \\ &\leq \mathbb{P}\left[\bigcup_{j=2}^{M} \{\imath(\mathbf{C}_1(n); Y^n) \leq \imath(\mathbf{C}_j(n); Y^n)\} \middle| W = 1\right] \,, && (151) \end{aligned}$$

where (150) follows from symmetry and (151) simply states that if $\tau > n$ and $W = 1$ then at least one information density should not be smaller than $\imath(\mathbf{C}_1(n); Y^n)$. We can proceed from (151) as in the random-coding union (RCU) bound [12, Theorem 17]:

$$\mathbb{P}[\tau > n] \leq \mathbb{E}\left[\min\left\{1, (M-1)\mathbb{P}[\imath(X^n; Y^n) \leq \imath(\bar{X}^n; Y^n) | X^n Y^n]\right\}\right] , \tag{152}$$

where we have additionally noted that conditioned on $W=1$ the joint distribution of $(\mathbf{C}_1(n), \mathbf{C}_j(n), Y^n)$ coincides with that of $(X^n, \bar{X}^n, Y^n)$ for every $j \neq 1$. Summing (152) over all n from 0 to ∞ we obtain

$$\mathbb{E}[\tau] = \sum_{n=0}^{\infty} \mathbb{P}[\tau > n] \leq \sum_{n=0}^{\infty} \mathbb{E}\left[\min\left\{1, (M-1)\mathbb{P}[\imath(X^n;Y^n) \leq \imath(\bar{X}^n;Y^n)|X^nY^n]\right\}\right]. \quad (153)$$

Thus, there must exist a realization of $\{\mathbf{C}_j, j=1,\ldots,M\}$ achieving (145). ■

Theorem 11: For an arbitrary DMC we have

$$\log M_{\mathsf{t}}^*(\ell, 0) = \ell C + O(\log \ell). \quad (154)$$

More specifically we have

$$\log M_{\mathsf{t}}^*(\ell, 0) \leq \ell C + \log \ell + O(1), \quad (155)$$

$$\log M_{\mathsf{t}}^*(\ell, 0) \geq \ell C + O(1). \quad (156)$$

Furthermore, the encoder achieving (156) uses feedback to calculate the stopping time only, i.e. it is an FV code.

Proof: The upper bound (155) follows from (48). To prove a lower bound, we will apply Theorem 10 with the process X selected as i.i.d. with a capacity-achieving marginal distribution. We first weaken the bound (145) to a form that is easier to analyze:

$$\mathbb{E}\left[\min\left\{1, (M-1)\mathbb{P}[\imath(X^n;Y^n) \leq \imath(\bar{X}^n;Y^n)|X^nY^n]\right\}\right] \quad (157)$$

$$\leq \mathbb{E}\left[\min\left\{1, M\mathbb{P}[\imath(X^n;Y^n) \leq \imath(\bar{X}^n;Y^n)|X^nY^n]\right\}\right] \quad (158)$$

$$= \mathbb{E}\left[\min\left\{1, M\mathbb{P}[\imath(X^n;Y^n) \leq \imath(\bar{X}^n;Y^n)|X^nY^n]\right\} 1\{\imath(X^n;Y^n) \leq \log M\}\right]$$
$$+ \mathbb{E}\left[\min\left\{1, M\mathbb{P}[\imath(X^n;Y^n) \leq \imath(\bar{X}^n;Y^n)|X^nY^n]\right\} 1\{\imath(X^n;Y^n) > \log M\}\right] \quad (159)$$

$$\leq \mathbb{P}[\imath(X^n;Y^n) \leq \log M] + M\mathbb{P}[\imath(\bar{X}^n;Y^n) > \log M] \quad (160)$$

$$= \mathbb{E}\left[\exp\left\{-[\imath(X^n;Y^n) - \log M]^+\right\}\right], \quad (161)$$

where (160) is obtained from (159) by upper-bounding $\min$ by 1 in the first term and by $M\mathbb{P}[\imath(\bar{X}^n;Y^n) > \log M]$ in the second term, and (161) is an application of (110).

In view of (161), Theorem 10 guarantees the existence of an $(\ell, M, 0)$ VLFT code with[8]

$$\ell \leq \mathbb{E}\left[\sum_{n=0}^{\infty} \exp\left\{-[\imath(X^n;Y^n) - \log M]^+\right\}\right]. \quad (162)$$

[8] $\imath(X^0;Y^0) = 0$ by convention.

We now define the filtration $\mathscr{F}$ as

$$\mathscr{F}_n = \sigma\{X^n, \bar{X}^n, Y^n\}, \quad n = 0, 1, \dots \tag{163}$$

Notice that $\imath(X^n; Y^n)$ is a random walk adapted to $\mathscr{F}$ with bounded jumps and positive drift equal to the capacity C:

$$\mathbb{E}\left[\imath(X^n; Y^n)\right] = nC\,, \tag{164}$$

whereas the process $\imath(\bar{X}^n; Y^n)$ is also a random walk with bounded jumps, but with a negative drift equal to the lautum information [22]:

$$\mathbb{E}\left[\imath(\bar{X}^n; Y^n)\right] = -nD(P_X P_Y || P_{XY}) = -nL(X; Y)\,. \tag{165}$$

Define a stopping time of the filtration $\mathscr{F}$ as follows:

$$\tau = \inf\{n \geq 0 : \imath(X^n; Y^n) \geq \log M\}\,. \tag{166}$$

With this definition we have

$$\mathbb{E}\left[\sum_{n=0}^{\infty} \exp\left\{-[\imath(X^n; Y^n) - \log M]^+\right\}\right] = \mathbb{E}\left[\tau + \sum_{k=0}^{\infty} \exp\left\{-[\imath(X^{k+\tau}; Y^{k+\tau}) - \log M]^+\right\}\right]. \tag{167}$$

Because $\imath(X^\tau; Y^\tau) \geq \log M$ we have

$$\begin{aligned} [\imath(X^{k+\tau}; Y^{k+\tau}) - \log M]^+ &= [\imath(X^{k+\tau}; Y^{k+\tau}) - \imath(X^\tau; Y^\tau) + \imath(X^\tau; Y^\tau) - \log M]^+ & (168)\\ &\geq [\imath(X^{k+\tau}; Y^{k+\tau}) - \imath(X^\tau; Y^\tau)]^+\,. & (169)\end{aligned}$$

Application of (169) gives

$$\mathbb{E}\left[\sum_{k=0}^{\infty} \exp\left\{-[\imath(X^{k+\tau}; Y^{k+\tau}) - \log M]^+\right\}\right] \leq \mathbb{E}\left[\sum_{k=0}^{\infty} \exp\left\{-[\imath(X^{k+\tau}; Y^{k+\tau}) - \imath(X^\tau; Y^\tau)]^+\right\}\right]. \tag{170}$$

By the strong Markov property of the random walk, conditioned on $\mathscr{F}_\tau$ the distribution of the process $\imath(X^{k+\tau}; Y^{k+\tau}) - \imath(X^\tau; Y^\tau)$ is the same as that of the process $\imath(X^k; Y^k)$. Thus, (167) and (170) imply

$$\mathbb{E}\left[\sum_{n=0}^{\infty} \exp\left\{-[\imath(X^n; Y^n) - \log M]^+\right\}\right] \leq \mathbb{E}\left[\tau\right] + \mathbb{E}\left[\sum_{k=0}^{\infty} \exp\left\{-[\imath(X^k; Y^k)]^+\right\}\right]. \tag{171}$$

To estimate the second term, notice that for some constants $a_1, a_2 > 0$ we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-[\imath(X^k;Y^k)]^+\right\}\right] && (172)\\
&= \mathbb{P}[\imath(X^k;Y^k)\le 0] + \mathbb{E}\left[\exp\left\{-\imath(X^k;Y^k)\right\}1\{\imath(X^k;Y^k>0\}\right] && (173)\\
&= \mathbb{P}[\imath(X^k;Y^k)\le 0] + \mathbb{P}[\imath(\bar{X}^k;Y^k)>0] && (174)\\
&\le a_2\exp\{-a_1 k\}\,, && (175)
\end{aligned}$$

where (174) is an application of (110), and (175) follows from Chernoff bound since both $\imath(X^k;Y^k)$ and $\imath(\bar{X}^k;Y^k)$ are sums of k i.i.d. random variables with positive expectation C and negative expectation $L(X;Y)$, respectively. Summing (175) over all non-negative integers k we obtain that for some constant $a_3 > 0$ we have

$$\mathbb{E}\left[\sum_{k=0}^{\infty}\exp\left\{-[\imath(X^k;Y^k)]^+\right\}\right] \le a_3\,. \tag{176}$$

Finally, by the boundedness of jumps of $\imath(X^n;Y^n)$ there is a constant $a_4 > 0$ such that

$$\imath(X^\tau;Y^\tau) - \log M \le a_4\,. \tag{177}$$

Since $\imath(X^n;Y^n) - nC$ is a martingale with bounded increments we have from Doob's stopping time theorem [17, Theorem 10.10]:

$$\mathbb{E}\left[\imath(X^\tau;Y^\tau)\right] = C\,\mathbb{E}\left[\tau\right], \tag{178}$$

but, on the other hand, from (177) we have

$$\mathbb{E}\left[\imath(X^\tau;Y^\tau)\right] \le \log M + a_4 \tag{179}$$

and thus,

$$\mathbb{E}\left[\tau\right] \le \frac{\log M}{C} + a_4\,. \tag{180}$$

Together (180), (176) imply via (171) and (162) the required lower bound (156). ■

Theorem 11 suggests that VLFT codes may achieve capacity even at very short blocklengths. To illustrate this numerically we first notice that Theorem 10 particularized to the BSC with i.i.d. input process X and an equiprobable marginal distribution yields the following result[9].

[9]This expression is to be compared with the (almost) optimal non-feedback achievability bound for the BSC, [12, Theorem 34].

Corollary 12: For the BSC with crossover probability δ and for every positive integer M there exists an $(\ell, M, 0)$ VLFT code satisfying

$$\ell \le \sum_{n=0}^{\infty}\sum_{t=0}^{n}\binom{n}{t}\delta^t(1-\delta)^{n-t}\min\left\{1,\ M\sum_{k=0}^{t}\binom{n}{k}2^{-n}\right\}. \tag{181}$$

A comparison of (181) and the upper bound (48) is given in Fig. 3. We see that despite the requirement of zero probability of error, VLFT codes are able to attain the capacity of the BSC at blocklengths as short as 30. As in Theorem 7 the convergence to capacity is very fast. Additionally, we have depicted the (approximate) performance of the best non-feedback code paired with the simple ARQ strategy, see [12, Section IV.E]. Note that the ARQ strategy indeed gives a valid zero-error VLFT code. The comparison on Fig. 3 suggests that even having access to the best possible block codes the ARQ is considerably suboptimal. It is interesting to note in this regard, that a Yamamoto-Itoh [5] strategy also pairs the best block code with a noisy version of ARQ (therefore, it is a VLF achievability bound). Consequently, we expect a similar gap in performance.

Another property of VLFT codes is that the maximal achievable rate for very small blocklengths may be noticeably above capacity. This can be seen as an artifact of the model which provides for an error-free termination symbol. Ordinarily, the overhead required in a higher layer to provide much higher reliability than the individual physical-layer symbols would not make short blocklengths attractive. This point is best demonstrated by computing the following specialized achievability bound for the BEC, which improves the general Theorem 10 in this particular case.

Theorem 13: For the BEC with erasure probability δ and any positive integer M there exists an $(\mu(M), M, 0)$ VLFT code, where the function $\mu : \mathbb{Z}_+ \to \mathbb{R}_+$ is the solution to

$$\begin{aligned}\mu(M) &= \frac{M-1}{M} + \delta\cdot\frac{1}{M}(M-1)\mu(M-1)\\ &\quad + (1-\delta)\cdot\frac{1}{M}\left[\left\lceil\frac{M-1}{2}\right\rceil\mu\left(\left\lceil\frac{M-1}{2}\right\rceil\right) + \left\lfloor\frac{M-1}{2}\right\rfloor\mu\left(\left\lfloor\frac{M-1}{2}\right\rfloor\right)\right]\end{aligned} \tag{182}$$

initialized by $\mu(1) = 0$.

Proof: If we need to transmit only one message, $M = 1$, then we can simply set $\tau = 0$. Therefore, we have

$$\mu(1) = 0\,. \tag{183}$$

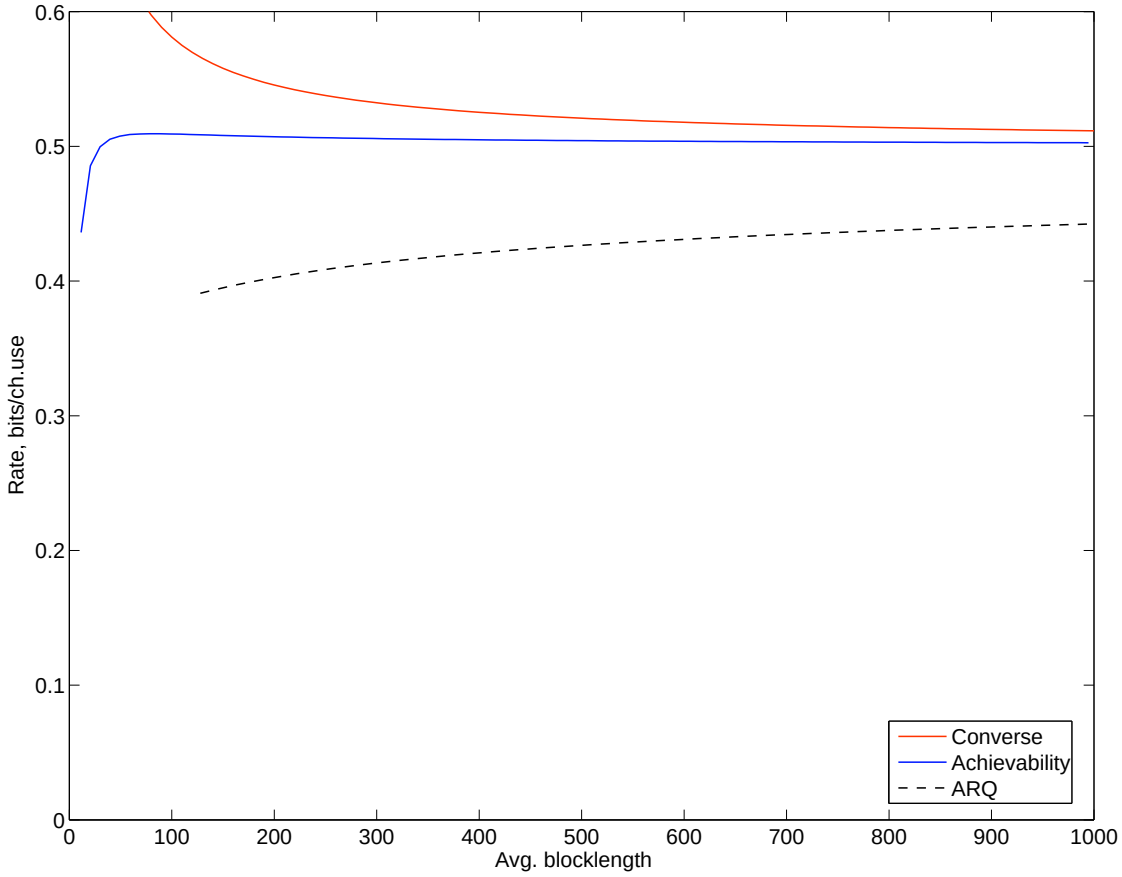

Fig. 3. Rate $\frac{1}{\ell}\log M_t^*(\ell,0)$ as a function of ℓ for zero-error transmission over the BSC(0.11) with a termination symbol. The lower bound is (181); the upper-bound is (48).

If we need to transmit an arbitrary $M>1$ number of messages than we do the following. First, all M messages are split into three groups. This splitting is static and known to both the encoder and the decoder. The first group consists of a single message ("a special message" below) and the remaining $M-1$ messages are split almost evenly in two ("non-special") groups, according to

$$M-1=\left\lceil\frac{M-1}{2}\right\rceil+\left\lfloor\frac{M-1}{2}\right\rfloor. \tag{184}$$

Second, if W is equal to the special message, then the encoder terminates the communication by setting $\tau=0$. If W belongs to one of $\left\lceil\frac{M-1}{2}\right\rceil$ messages then the encoder sets $f_1=0$, and to $f_1=1$ if W belongs to the remaining group of $\left\lfloor\frac{M-1}{2}\right\rfloor$. Third, upon passing through the channel one of three possibilities can happen: transmission terminated with T (if W was a

special message), the digit was delivered correctly, or the digit was erased:

1) In the first case, the decoder knows that W must have been equal to the pre-selected special message, which it outputs as $\hat{W}$ (error-free, of course).
2) In the second case the decoder has gained the knowledge to which of the two non-special groups W belonged. Therefore, we can reiterate the algorithm with a reduced size of the message set, setting $M' = \lceil \frac{M-1}{2} \rceil$ or $M' = \lfloor \frac{M-1}{2} \rfloor$, depending on the group to which W belonged.
3) Finally, if the digit was erased then the only knowledge the decoder has gained is that W was a non-special message. Therefore, we reiterate the algorithm with $M' = M - 1$ since the special message was ruled out.

We now analyze the average number of channel uses required for such a recursive procedure to terminate. The first case happens with probability $\frac{1}{M}$ and the total number of channel uses is 0. The second case happens with probability $\frac{M-1}{M} \cdot (1-\delta)$ and the (average) number of channel uses is $1+\mu\left(\lceil \frac{M-1}{2} \rceil\right)$ or $1+\mu\left(\lfloor \frac{M-1}{2} \rfloor\right)$ depending on the group to which W belonged. Finally, the third case happens with probability $\frac{M-1}{M} \cdot \delta$ and the number of channel uses is $1+\mu(M-1)$. In total we obtain (182). ■

The first few values of the μ-function are

$$\mu(1) = 0\,, \tag{185}$$

$$\mu(2) = 1/2\,, \tag{186}$$

$$\mu(3) = \frac{1}{3}(2+\delta)\,, \tag{187}$$

$$\mu(4) = 1+\frac{1}{4}(\delta+\delta^2)\,. \tag{188}$$

Since it is not possible to compute $\mu(2^{500})$ directly, the following idea was used for large values of M. Fix some k_{max} and compute $\mu(2^k)$ for all $k \le k_{max}$ via (182). For $k > k_{max}$ we can use a strategy of simply retransmitting each of the first $k - k_{max}$ bits until it is delivered unerased, and then use the described recursive strategy to transmit the remaining k_{max} bits. This gives the bound

$$\mu(2^k) \le \frac{k-k_{max}}{1-\delta} + \mu\left(2^{k_{max}}\right)\,. \tag{189}$$

As k_{max} increases, the upper bound improves. Experimentation shows that there is no visible improvement once $k_{max} \gtrsim 10$.

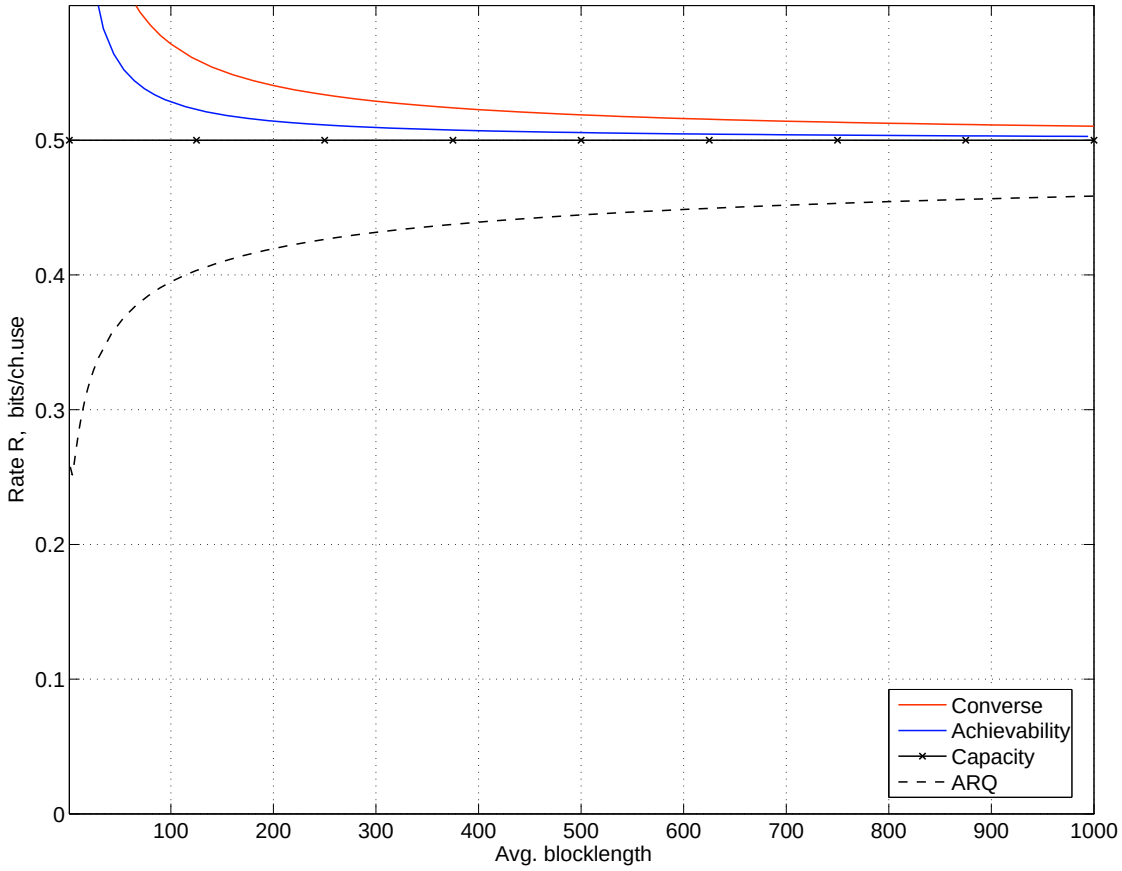

Fig. 4. Rate $\frac{1}{\ell}\log M_t^*(\ell,0)$ as a function of ℓ for zero-error transmission over the BEC(0.5) with a termination symbol. The lower bound is Theorem 13; the upper-bound is (48).

Numerical comparison of the achievability bound of Theorem 13 against the converse bound (48) is given on Fig. 4 for the case of $\delta = 0.5$. We notice that indeed for small ℓ (and, equivalently, M) the availability of the termination symbol allows the rate to exceed the capacity slightly. Also, the horizontal capacity line coincides with the "traditional" achievability bound for the BEC, as given by Theorem 7 with $\epsilon = 0$, which does not take advantage of the additional degree of freedom enabled in the VLFT paradigm (i.e., a termination symbol).

V. Excess Delay Constraints

Quantifying the notion of delay for variable-length coding with feedback has proven to be notoriously hard, see, for example, [23] for a related discussion. While for fixed-blocklength codes, delay is naturally identified with blocklength, in the variable-length setup, however,

the usage of average blocklength $\mathbb{E}[\tau]$ as a proxy for delay may not be sensible in real-time applications with hard delay constraints. Nevertheless, the definition of rate as $\frac{\log M}{\mathbb{E}[\tau]}$ is very natural, since by the law of large numbers, the ratio of bits to channel uses will approach $\frac{\log M}{\mathbb{E}[\tau]}$ for a repeated use of the same code.

An advantage of feedback is the ability to terminate transmission early on favorable noise realizations thereby reducing average blocklength. However, it remains to be seen whether under a constraint on the probability of excess delay, variable-length coding offers any advantage. Depending on whether VLF or VLFT codes are used, we consider two different formulations of the delay constraint. While the delay is not allowed to exceed a certain threshold in either case, for the VLFT codes we additionally require the decoded message to be correct with probability one.

A. VLF codes

Consider an arbitrary VLF code and define the error event differently from (6). Namely, fix a delay d and define the probability of error as

$$\epsilon = \mathbb{P}[\{\hat{W} \neq W\} \cup \{\tau > d\}]\,. \tag{190}$$

The question is then: what is the maximum M compatible with a chosen d and ϵ? The answer is obvious: since in such formulation the encoder has no incentive to terminate before the delay d, we might as well fix blocklength to be d and force the decoder to take the decision at time d. This, however, is no different from fixed-blocklength coding with feedback.

Definition 3: An (n, M, ϵ) fixed-blocklength feedback code is an (n, M, ϵ) VLF code with $\tau = n$. The fundamental limit of fixed-blocklength feedback codes is given by

$$M_b^*(n, \epsilon) = \max\{M : \exists (n, M, \epsilon) \text{ fixed-length feedback code}\}\,. \tag{191}$$

In the case of the BEC, the tight converse bound for fixed-blocklength codes shown in [12, Theorem 38] applies even when feedback is available. Therefore, the proof of [12, Theorem 53] automatically applies to the feedback case and we have:

Theorem 14: For the BEC,

$$\log M_b^*(n, \epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(1)\,, \tag{192}$$

where C and V are the capacity and the dispersion of the BEC.

Therefore, we see that if we treat the excess delay as error, see (190), then feedback is unable to improve the $\sqrt{n}$ penalty term. In fact, much more is true. The numerical computation of the upper (converse) bound for the BEC was shown in [12, Section III.I] to be extremely tight. In particular, it was shown that non-feedback codes exist that achieve values of $\log_2 M$ within 2-3 bits of the converse bound for all blocklengths $n \gtrsim 10$. Consequently, under an excess delay probability constraint, the potential benefit of feedback is limited to enlarging $\log_2 M$ by those 2-3 bits at most.

Similar conclusions regarding the expansion and the bounds hold for a wide class of symmetric channels (including the BSC), as is shown below. Namely, we demonstrate that for such channels, the expansion (21) does not change in the presence of feedback when attention is restricted to the fixed-blocklength codes defined in Definition 3. Moreover, we demonstrate that non-asymptotic (converse) upper bounds, used for numerical computations in [12, Sections III.H, III.I] and shown there to be extremely tight, hold verbatim in the presence of feedback. The main idea in our analysis is due to Dobrushin [2] and thus, the subsequent results may be viewed as both a strengthening of his result to the non-asymptotic setting of [12], and as a generalization to a wider class of channels defined as follows.

Definition 4: A DMC $(\mathcal{A}, \mathcal{B}, P_{Y|X})$ is called weakly input-symmetric if there exists an $x_0 \in \mathcal{A}$ and a random transformation $T_x : \mathcal{B} \to \mathcal{B}$ for each $x \in \mathcal{A}$ such that $T_x \circ P_{Y|X=x_0} = P_{Y|X=x}$ and $T_x \circ P_{Y^*} = P_{Y^*}$, where P_{Y^*} is the (unique) capacity achieving output distribution.
Note that the composition $T_x \circ P_Y$ with a distribution P_Y on $\mathcal{B}$ is given by

$$(T_x \circ P_Y)(y) = \sum_{y' \in \mathcal{B}} T_x(y|y') P_Y(y') \,. \tag{193}$$

Thus, in other words, T_x is a stochastic matrix which upon multiplication by the column $P_{Y|X=x_0}$ yields the column $P_{Y|X=x}$. Weak input-symmetry is a (strict) generalization of Dobrushin [2] and Gallager [24, p. 94] symmetries. This more general property, however, is sufficient to compute the $\log n$ term in (21); see [16, Section 3.4.5].

The performance of an optimal binary hypothesis test is defined as follows (see [12, Section III.D2] for more details). Consider a W-valued random variable W that can take probability measures P or Q. A randomized test between those two distributions is defined by a random transformation $P_{Z|W} : \mathsf{W} \mapsto \{0, 1\}$ where 0 indicates that the test chooses Q. The best perfor-

mance achievable among those randomized tests is given by

$$\beta_\alpha(P,Q) = \min \sum_{w\in\mathsf{W}} Q(w)P_{Z|W}(1|w)\,, \tag{194}$$

where the minimum is over all probability distributions $P_{Z|W}$ satisfying

$$\sum_{w\in\mathsf{W}} P(w)P_{Z|W}(1|w) \geq \alpha\,. \tag{195}$$

The minimum in (194) is guaranteed to be achieved by the Neyman-Pearson lemma. Thus, $\beta_\alpha(P,Q)$ gives the minimum probability of error under hypothesis Q if the probability of error under hypothesis P is not larger than $1-\alpha$.

The generalization of Theorem 14 is the following:

Theorem 15: Consider a weakly input-symmetric DMC with capacity C and dispersion V. Then $M_b^*(n,\epsilon)$ satisfies a "sphere-packing bound":

$$M_b^*(n,\epsilon) \leq \frac{1}{\beta_{1-\epsilon}^n}\,, \tag{196}$$

where β_α^n is defined for $0\leq\alpha\leq 1$ as follows:

$$\beta_\alpha^n \stackrel{\triangle}{=} \beta_\alpha(P_{Y|X=x_0}^n, P_{Y^*}^n)\,, \tag{197}$$

with $x_0\in\mathcal{A}$ and P_{Y^*} being as defined in Definition 4. In particular, if $V>0$ then as $n\to\infty$ we have

$$\log M_b^*(n,\epsilon) \leq nC - \sqrt{nV}Q^{-1}(\epsilon) + \frac{1}{2}\log n + O(1)\,, \tag{198}$$

whereas if $V=0$ then

$$\log M_b^*(n,\epsilon) \leq nC - \log(1-\epsilon)\,. \tag{199}$$

For example, for the BSC it was shown in [12, Section III.H] that (196) is tight to within a factor of 10 for a wide range of n. Therefore, for the BSC and such n, the value of $\log M_b^*(n,\epsilon)$ can improve the $\log M^*(n,\epsilon)$ (achieved without feedback) by at most 3-4 bits.

Some properties of weakly input-symmetric channels (for notation see [12, Section IV.A]) are summarized in the next result.

Theorem 16: For any weakly input-symmetric DMC W all of the following hold:

1) The capacity C satisfies

$$C = D(P_{Y|X=x_0}||P_{Y^*})\,. \tag{200}$$

2) The ϵ-dispersion V_ϵ, see [12, Definition 2], equals the dispersion V and satisfies

$$V = V(P_{Y|X=x_0}||P_{Y^*}) \tag{201}$$

$$= V(P_{Y|X=x}||P_{Y^*}) \quad (\forall x : D(P_{Y|X=x}||P_{Y^*}) = C)\,. \tag{202}$$

3) If $V > 0$ then as $n \to \infty$ we have

$$-\log \beta^n_{1-\epsilon} = nC - \sqrt{nV}Q^{-1}(\epsilon) + \frac{1}{2}\log n + O(1)\,. \tag{203}$$

If $V = 0$ then we have

$$-\log \beta^n_{1-\epsilon} = nC - \log(1-\epsilon)\,. \tag{204}$$

Proof: To show (200) notice that a transformation T_x maps the pair of distributions $(P_{Y|X=x_0}, P_{Y^*})$ to $(P_{Y|X=x}, P_{Y^*})$ and therefore by the data processing for divergence we get

$$D(P_{Y|X=x}||P_{Y^*}) \le D(P_{Y|X=x_0}||P_{Y^*})\,, \tag{205}$$

from which (200) follows via

$$C = \max_{x\in\mathcal{A}} D(P_{Y|X=x}||P_{Y^*})\,. \tag{206}$$

For each x^n define a random transformation $T_{x^n} : \mathcal{B}^n \to \mathcal{B}^n$ as follows:

$$T_{x^n}(z^n|y^n) = \prod_{k=1}^n T_{x_k}(z_k|y_k)\,. \tag{207}$$

Then T_{x^n} maps the pair of distributions $(P^n_{Y|X=x_0}, P^n_{Y^*})$ to $(P_{Y^n|X^n=x^n}, P^n_{Y^*}$ and thus by the data-processing property for β_α (i.e., application of a random transformation cannot improve the value of β_α) we obtain

$$\beta_\alpha(P_{Y^n|X^n=x^n}, P^n_{Y^*}) \ge \beta_\alpha((P^n_{Y|X=x_0}), P^n_{Y^*})\,. \tag{208}$$

To show (201) notice that by [12, Lemma 58] we have for any $x \in \mathcal{A}$,

$$\log \beta_\alpha(P^n_{Y|X=x}, P^n_{Y^*}) = -nD(P_{Y|X=x}||P_{Y^*}) - \sqrt{nV(P_{Y|X=x}||P_{Y^*})}Q^{-1}(\alpha) + o(\sqrt{n})\,. \tag{209}$$

But by (208) we must have

$$\log \beta_\alpha(P^n_{Y|X=x}, P^n_{Y^*}) \ge \log \beta_\alpha(P^n_{Y|X=x_0}, P^n_{Y^*})\,. \tag{210}$$

Now assuming that $x \in \mathcal{A}$ is such that $D(P_{Y|X=x}||P_{Y^*}) = C$ and applying (209) to both sides of (210) for $\alpha > 1/2$ we obtain

$$V(P_{Y|X=x}||P_{Y^*}) \ge V(P_{Y|X=x_0}||P_{Y^*})\,, \tag{211}$$

whereas taking $\alpha < 1/2$ we show

$$V(P_{Y|X=x}||P_{Y^*}) \le V(P_{Y|X=x_0}||P_{Y^*})\,, \tag{212}$$

and consequently (201) follows.

Finally, by (200), (201) and [12, Lemma 58] (see also [16, (2.89)-(2.90)]) we obtain (203) and (204). ■

Proof of Theorem 15: Fix an (n, M, ϵ) fixed-blocklength feedback code. Its encoder defines a transition probability kernel $P_{Y^n|W}$ from the input space

$$\mathcal{D}_M \stackrel{\triangle}{=} \{1, \ldots, M\} \tag{213}$$

to the output space $\mathcal{B}^n$. We can view then the triplet $(\mathcal{D}_M, P_{Y^n|W}, \mathcal{B}^n)$ as a channel on which we have a usual (M, ϵ) code. For such a code [12, Theorem 27] shows

$$M \le \frac{1}{\beta_{1-\epsilon}(P_{WY^n}, P_W Q_{Y^n})}\,, \tag{214}$$

where P_W is the equiprobable distribution on $\mathcal{D}_M$ and Q_{Y^n} is the following product distribution on $\mathcal{B}^n$:

$$Q_{Y^n}(y_j) = \prod_{j=1}^{n} P_{Y^*}(y_j)\,. \tag{215}$$

Therefore, the proof of (196) will be complete if we can show

$$\beta_\alpha(P_{WY^n}, P_W Q_{Y^n}) \ge \beta_\alpha^n\,. \tag{216}$$

Lemma 17 (at the end of this section) shows that (216) follows if we prove that for any $j \in \{1, \ldots, M\}$

$$\beta_\alpha(P_{Y^n|W=j}, Q_{Y^n}) \ge \beta_\alpha^n\,. \tag{217}$$

Fix arbitrary $j \in \{1, \ldots, M\}$ and $x_0 \in \mathcal{A}$. The sequence of encoder functions $f_k, k = 1, \ldots, n$ defines the measure $P_{Y^n|W=j}$ as follows:

$$P_{Y^n|W=j}(y^n) = \prod_{k=1}^{n} P_{Y|X}(y_k|f_k(j, y^{k-1}))\,. \tag{218}$$

Since the channel is weakly input-symmetric, to each $x \in \mathcal{A}$ there exists a transformation $T_x : \mathcal{B} \to \mathcal{B}$ such that

$$P_{Y|X=x} = T_x \circ P_{Y|X=x_0}\,, \tag{219}$$

where the composition is understood as in (193). We will now define a transformation $T_j : \mathcal{B}^n \to \mathcal{B}^n$ as follows:

$$T_j(z^n|y^n) = \prod_{k=1}^{n} T_{f_k(j,y^{k-1})}(z_k|y_k)\,. \tag{220}$$

Then according to this construction and (218), on the one hand we have

$$T_j \circ P^n_{Y|X=x_0} = P_{Y^n|W=j}\,, \tag{221}$$

whereas on the other hand, since each T_x preserves P_{Y^*}, we also have

$$T_j \circ Q_{Y^n} = Q_{Y^n}\,. \tag{222}$$

Then it follows that

$$\begin{aligned} \beta_\alpha(P_{Y^n|W=j}, Q_{Y^n}) &= \beta_\alpha(T_j \circ P^n_{Y|X=x_0}, T_j \circ Q_{Y^n}) && (223)\\ &\geq \beta_\alpha(P^n_{Y|X=x_0}, Q_{Y^n})\,, && (224)\end{aligned}$$

where (223) follows by (221) and (222), and (224) follows by data-processing property for β_α (i.e., simultaneous application of T_j to both measures cannot improve the value of β_α). This completes the proof of (196). Finally, (198) and (199) follow by (214) and (203) or (204), respectively. ■

The following result is a straightforward modification of [12, Lemma 29]; see [16, Lemma 32]:

Lemma 17: Suppose that there is an non-decreasing convex function $f : [0,1] \to [0,1]$ such that for all $x \in \mathsf{F}$

$$\beta_\alpha(P_{Y|X=x}, Q_{Y|X=x}) \geq f(\alpha)\,. \tag{225}$$

Then, for any P_X supported on F we have

$$\beta_\alpha(P_X P_{Y|X}, P_X Q_{Y|X}) \geq f(\alpha)\,. \tag{226}$$

B. VLFT codes

It has been shown above that one of the key advantages of VLFT codes is in their ability to achieve zero probability of error without any penalty in rate. In this section we study the question of excess delay for such codes. For a zero-error VLFT code we define an ϵ-delay as

$$D_\epsilon = \min\{n : \mathbb{P}[\tau > n] \leq \epsilon\}\,, \quad \epsilon \in [0,1]. \tag{227}$$

Thus a zero-error VLFT code with $D_\epsilon \le d$ is a code which is guaranteed to deliver the data error-free, and does so in less than d channel uses in all except ϵ-portion of the cases. The question arises: for a fixed ϵ, what is the maximum M compatible with a given ϵ-delay requirement d:

$$M_z^*(d,\epsilon) = \max\{M : \exists \text{zero-error VLFT code with } D_\epsilon \le d\}\,? \tag{228}$$

The obvious achievability bound is to simply pair a fixed-blocklength non-feedback (n, M, ϵ) with $n = d$ code with an ARQ retransmission strategy to achieve zero error. We have thus

$$M_z^*(d,\epsilon) \ge M^*(d,\epsilon) = dC - \sqrt{dV}Q^{-1}(\epsilon) + O(\log d)\,, \tag{229}$$

where $M^*(d,\epsilon)$ denotes the performance of the best non-feedback, fixed-blocklength code and is thus given by (21).

Can we improve the crucial $\sqrt{d}$-penalty term in (229)? The answer is negative, at least for the BEC:

Theorem 18: For the BEC,

$$\log M_z^*(d,\epsilon) \le dC - \sqrt{dV}Q^{-1}(\epsilon) + \log d + O(1)\,, \tag{230}$$

where C and V are the capacity and the dispersion of the BEC.

Proof: Let E_j be the i.i.d. process corresponding to erasures: $\mathbb{P}[E_j = 0] = 1 - \mathbb{P}[E_j = 1] = \delta$, where δ is the erasure probability of the BEC. Then the total number of unerased symbols by time n is given by

$$N_n = \sum_{j=1}^{n} E_j\,. \tag{231}$$

Following the steps of the proof of [12, Theorem 38], we can see that by time n the total number of messages distinguishable at the receiver is upper-bounded by $\sum_{j=0}^{n} 2^{N_j}$ (summation corresponds to the fact that a VLFT code has the freedom of sending a termination symbol at any time). Therefore, since the code achieves zero-error we have

$$\mathbb{P}[\tau \le n] \le \frac{1}{M}\mathbb{E}\left[\min\left\{\sum_{j=0}^{n} 2^{N_j}\,, M\right\}\right]\,. \tag{232}$$

Since N_t is a monotonically non-decreasing it follows that

$$\begin{aligned}
\sum_{j=0}^{n} 2^{N_j} &\le \sum_{t=0}^{N_n} 2^t + (n - N_n)2^{N_n} && (233)\\
&\le 2^{N_n}(n + 2 - N_n) && (234)\\
&\le (n+2)2^{N_n}\,. && (235)
\end{aligned}$$

Although the bound (234) is useful for numerical evaluation, the bound (235) is more convenient for the analysis. Indeed, we have from (232) and (235):

$$\begin{aligned} \mathbb{P}[\tau \leq n] &\leq \frac{1}{M}\mathbb{E}\left[\min\left\{(n+2)2^{N_n}, M\right\}\right] && (236)\\ &= \frac{n+2}{M}\mathbb{E}\left[\min\left\{2^{N_n}, \frac{M}{n+2}\right\}\right]. && (237)\end{aligned}$$

Recall now that for the non-feedback case [12, Theorem 38] can be restated as

$$1-\epsilon \leq \frac{1}{M}\mathbb{E}\left[\min\left\{2^{N_n}, M\right\}\right]. \tag{238}$$

The analysis of the bound (238) in the proof of [12, Theorem 53], has shown that (238) implies

$$\log M \leq nC - \sqrt{nV}Q^{-1}(\epsilon) + O(1), \tag{239}$$

as $n \to \infty$, where C and V are the capacity and the dispersion of the BEC. Comparing (238) and (237) we see that M is replaced by $\frac{M}{n+2}$. Therefore, the same argument as the one leading from (238) to (239) when applied to (237) must give

$$\log M \leq nC - \sqrt{nV}Q^{-1}(\epsilon) + \log(n+2) + O(1), \tag{240}$$

which implies (230). ■

VI. Discussion

We have demonstrated that by allowing variable length, even a modicum of feedback is enough to considerably speed up convergence to capacity. For illustration purposes we can see in Fig. 2 that we have constructed a stop-feedback code, that achieves, for example, 90% of the capacity of the BSC with crossover probability $\delta = 0.11$ and probability of error $\epsilon = 10^{-3}$ at blocklength 200; see Fig. 2. In contrast, to obtain the same performance with fixed-blocklength codes requires a blocklength of at least 3100 even if full noiseless feedback is available at the transmitter. This practical benefit of feedback opens the possibility of utilizing the full capacity of the link without the complexity required to implement coding of very long data packets.

A major ingredient of the achievability bounds in this paper is the idea of terminating early on favorable noise realizations. Although, we have applied this idea to the codes with codewords with unbounded durations, it is clear that without any significant effect on probability of error we could also assume that the transmission forcibly terminates after a time which is a few times the

average blocklength ℓ. Consequently, it can be shown that any point on the achievability curve of Fig. 2 can be realized by pairing some linear block code with the stopping rule (35). In other words, even traditional fixed-blocklength linear codes can be decoded with significantly less (average) delay if used in the variable-length setting. It is important, thus, to investigate whether traditionally good codes (such as low-density parity-check (LDPC) codes) are also competitive in this setting.

Theoretically, the benefit of feedback is manifested by the absence of the $\sqrt{\ell}$ term in the expansions (16) and (17), whereas this term is crucial to determine the non-asymptotic performance without feedback. Equivalently, we have demonstrated that for variable-length codes with feedback the channel dispersion is zero. To intuitively explain this phenomenon, we note that without feedback the main effect governing the $\sqrt{n}$ behavior was the stochastic variation of information density around its mean, which is tightly characterized by the central limit theorem. In the variable-length setup with feedback, the main idea is that of Wald-like stopping once the information density of some message is large enough. Therefore, there is virtually no stochastic variation (besides a negligible overshoot) and this explains the absence of any references to the central limit theorem.

We have also analyzed a modification of the coding problem by introducing a termination symbol (VLFT codes), which is motivated in many practical situations in which control signals are sent over a highly reliable upper layer. We have shown that in this setup, in addition to the absence of $\sqrt{\ell}$ term, the principal new effect is that the zero-error capacity increases to the full Shannon capacity of the channel. Although availability of a "use-once" termination symbol is immaterial asymptotically, the transient behavior is significantly improved. Analytically, this effect is predicted by the absence of not only the $\sqrt{\ell}$ term but also of the $\log \ell$ term in the achievability bound (156). Furthermore, our codes with termination have a particularly convenient structure: the encoder uses the feedback link only to choose the time when to stop the transmission (by sending the termination symbol), and otherwise simply sends a fixed message-dependent codeword. The codes with such structure have been called fixed-to-variable (FV), or fountain, codes in [19]. Thus, in short, we have demonstrated that fountain codes can achieve 90% of the capacity of the BSC with crossover probability $\delta = 0.11$ at average blocklength < 20 and with zero probability of error. Practically, of course, "zero-error" should be understood as the reliability being essentially the probability with which the termination symbol is correctly

detected.

Finally, we have discussed some questions regarding communication of real-time data. We have demonstrated that constraints on the excess delay nullify the advantage of feedback (and variable length), i.e. the improvement in performance of the best feedback code can be marginal at best compared to non-feedback, fixed-blocklength codes. This is in sharp contrast with the results of the previous sections.

APPENDIX

The next result shows that restriction on the cardinality of $\mathcal{U}$ in the Definitions 1 and 2 does not entail loss of generality.

Theorem 19: For any (ℓ, M, ϵ) VLFT code there exists an (ℓ, M, ϵ) VLFT code with $|\mathcal{U}| \leq 3$.

Proof: Denote by G_k the following subsets of $\mathbb{R}^2$:

$$G_k \stackrel{\triangle}{=} \{(\ell', \epsilon') : \exists(\ell', M, \epsilon')\text{-code with } |\mathcal{U}| \leq k\}\,, k = 1, 2, \ldots, \tag{241}$$

and

$$G_\infty \stackrel{\triangle}{=} \{(\ell', \epsilon') : \exists(\ell', M, \epsilon')\text{-code}\}\,. \tag{242}$$

Notice that G_∞ is a convex hull of G_1 since by taking a general code and conditioning on U we obtain a deterministic code. By Caratheodory's theorem we then know that $G_3 = G_\infty$. Since by assumption $(\ell, \epsilon) \in G_\infty$ then $(\ell, \epsilon) \in G_3$. ■

Proof of Theorem 1: Fix $\epsilon' < \epsilon$ and a large n. Then there exists a fixed-blocklength code without feedback with blocklength n, probability of error ϵ' and number of messages M satisfying:

$$\log M \geq nC + o(n)\,. \tag{243}$$

Consider the following variable-length code (without feedback): with probability $\frac{1-\epsilon}{1-\epsilon'}$ encoder sends a codeword of length n, otherwise it sends nothing. It is easy to see that the probability of decoding error is upper-bounded by ϵ whereas the average transmission time is equal to $\ell = \frac{1-\epsilon}{1-\epsilon'}n$, and therefore the average transmission rate is

$$R \stackrel{\triangle}{=} \frac{\log M}{\ell} \geq C\frac{1-\epsilon'}{1-\epsilon} + o(1)\,. \tag{244}$$

By taking the limit $n \to \infty$ we obtain

$$[\![C_\epsilon]\!] \geq C\frac{1-\epsilon'}{1-\epsilon}\,. \tag{245}$$

Since ϵ' is arbitrary we can achieve any rate close to $\frac{C}{1-\epsilon}$.

For the converse recall that a channel is said to satisfy strong converse if its fixed-blocklength no feedback fundamental limit $\log M^*(n,\epsilon)$ satisfies

$$\log M^*(n,\epsilon) = nC + o(n)\,, n \to \infty\,, \quad \forall \epsilon \in (0,1)\,. \tag{246}$$

Now, consider an (ℓ, M, ϵ) variable-length code. Define the following quantities for each $n \geq 0$ and $u \in \mathcal{U}$:

$$\epsilon(n,u) \;\; = \;\; \mathbb{P}[\hat{W} \neq W | \tau = n, U = u]\,, \tag{247}$$

which satisfy, of course,

$$\mathbb{E}\left[\epsilon(\tau, U)\right] \leq \epsilon\,. \tag{248}$$

Fix u and notice that conditioned on $U = u$, τ is a function of W, and therefore $M\mathbb{P}[\tau = n|U = u]$ is an integer. Then the condition $\tau = n$ defines an $(n, M\mathbb{P}[\tau = n|U = u], \epsilon(n,u))$ fixed blocklength subcode. Therefore, we have for each $n \geq 0$:

$$\mathbb{P}[\tau = n|U = u]M \leq M^*(n, \epsilon(n,u))\,. \tag{249}$$

We now fix arbitrary $N \geq 0$ and $\epsilon' > 0$ and sum (249) for all $n \leq N$ such that $\epsilon(n,u) \leq \epsilon'$:

$$\begin{aligned} M\mathbb{P}[\tau \leq N, \epsilon(\tau,u) \leq \epsilon' | U = u] \;\; &\leq \;\; \sum_{n=0}^{N} M^*(n, \epsilon(n,u)) 1\{\epsilon(n,u) \leq \epsilon'\}\,, && (250)\\ &\leq \;\; \sum_{n=0}^{N} M^*(n, \epsilon')\,, && (251)\\ &\leq \;\; NM^*(N, \epsilon')\,, && (252) \end{aligned}$$

where (251) follows since by definition $M^*(n,\epsilon)$ is a non-decreasing function of ϵ, and (252) follows because for a non-anticipatory channel $M^*(n,\epsilon)$ is also a non-decreasing function of n. By taking the expectation of (252) with respect to U we obtain

$$M\mathbb{P}[\tau \leq N, \epsilon(\tau,U) \leq \epsilon'] \leq NM^*(N,\epsilon')\,. \tag{253}$$

On the other hand, by the Chebyshev inequality we have

$$\begin{aligned} \mathbb{P}[\tau \leq N, \epsilon(\tau,U) \leq \epsilon'] \;\; &\geq \;\; 1 - \frac{\mathbb{E}\left[\tau\right]}{N} - \frac{\mathbb{E}\left[\epsilon(\tau,U)\right]}{\epsilon'} && (254)\\ &\geq \;\; 1 - \frac{\ell}{N} - \frac{\epsilon}{\epsilon'}\,. && (255) \end{aligned}$$

Finally, we choose $\epsilon' > \epsilon$ and take

$$N = \frac{\ell+1}{1-\epsilon/\epsilon'}. \tag{256}$$

Now from (253), (255) and (256) we obtain

$$\log M \leq \log M^*\left(\frac{\ell+1}{1-\epsilon/\epsilon'}, \epsilon'\right) + 2\log\frac{\ell+1}{1-\epsilon/\epsilon'} \tag{257}$$

$$= C\frac{\ell+1}{1-\epsilon/\epsilon'} + o(\ell), \tag{258}$$

where (258) follows from strong converse (246). Dividing both sides of (258) by ℓ we have proven that the rate of any (ℓ, M, ϵ) variable-length code must satisfy:

$$\frac{\log M}{\ell} \leq \frac{C}{1-\epsilon/\epsilon'} + o(1), \tag{259}$$

or in other words, for any $\epsilon' > \epsilon$ we have

$$[\![C_\epsilon]\!] \leq \frac{C}{1-\epsilon/\epsilon'}. \tag{260}$$

Taking $\epsilon' \to 1$ completes the proof. ■

Proof of Theorem 6: As in the proof of Theorem 4 it is sufficient to consider the case of codes with $|\mathcal{U}| = 1$. This follows because of convexity of the right-hand side of (98) in ϵ as explained in (92). Next, by replacing a stopping time τ with $\min\{\tau, N\}$, $N = 1, \ldots$ and including $\{\tau > N\}$ in the error event, we obtain a sequence of codes with probability of error $\epsilon_N \searrow \epsilon$ as $N \to \infty$. Since for each fixed ξ the argument of the supremum in (98) is continuous in ϵ, it is sufficient to prove (98) for codes with a bounded $\tau \leq N$ for some $N \geq 1$.

We consider a measurable space $\Omega = \{1, \ldots, M\} \times \mathcal{B}^\infty$, understood as (W, Y^∞) with filtration $\mathscr{G}_n$ as in Definition 1. Fixing a code we notice that encoder $\{f_n, n = 1, \ldots\}$ induces a distribution $\mathbb{P} = P_{WY^\infty}$ on Ω. Considering a stopping time τ_1 (to be specified later), we define an auxiliary measure $\mathbb{Q}$ on Ω as follows:

$$\mathbb{Q}[W = j] = \frac{1}{M}, \tag{261}$$

$$\mathbb{Q}[Y_n = y_n | Y^{n-1} = y^{n-1}, W = j] = P_Y^*(y_n) 1\{n \leq \tau_1\} \tag{262}$$

$$+ \mathbb{P}[Y_n = y_n | Y^{n-1} = y^{n-1}, W \neq j] 1\{n > \tau_1\} \tag{263}$$

where P_Y^* is the unique capacity achieving output distribution corresponding to a DMC $P_{Y|X}$. For convenience we denote $P^j[\cdot] = \mathbb{P}[\cdot|W = j]$ and $Q^j[\cdot] = \mathbb{Q}[\cdot|W = j]$ for $j = 1, \ldots, M$. Then

we have for any event B

$$Q^j[B|\mathscr{G}_{\tau_1}] = \frac{\mathbb{P}[B, W \neq j|\mathscr{G}_{\tau_1}]}{\mathbb{P}[W \neq j|\mathscr{G}_{\tau_1}]}\,. \tag{264}$$

Notice that since $\tau \leq N$ we may replace $\mathcal{B}^\infty$ with $\mathcal{B}^N$ thereby reducing to the case of a finite space Ω. Moreover, because $C_1 < \infty$ measures $\mathbb{P}$, P^j, $\mathbb{Q}$, $(P_Y^*)^N$ and the counting measure are all mutually absolutely continuous. This enables us to avoid adding specifiers "almost surely" and dealing with non-uniqueness of conditional expectations in (264) and below.

We define the following processes

$$\begin{aligned}
S_n &= \log \frac{P_{Y^n|W}(Y^n|W)}{Q_{Y^n|W}(Y^n|W)}\,, && (265)\\
R_n &= S_n - \min(n, \tau_1)C - |n - \tau_1|^+ C_1\,, && (266)\\
\eta_n &= \mathbb{P}[\hat{W} \neq W|\mathscr{G}_n]\,, && (267)\\
\pi_n(w) &= \mathbb{P}[W = w|\mathscr{G}_n]\,, && (268)\\
\pi_n^{\max} &= \max_w \pi_n(w)\,, && (269)\\
\hat{W}_n &= \operatorname*{argmax}_w \pi_n(w)\,, && (270)
\end{aligned}$$

Without loss of generality we can assume that our code satisfies

$$\begin{aligned}
\hat{W} &\stackrel{\triangle}{=} g(Y^\tau) = \hat{W}_\tau\,, && (271)\\
\pi_\tau^{\max} &= 1 - \eta_\tau\,, && (272)\\
\pi_n^{\max} &\leq 1 - \eta_n\,, \qquad \forall\, 0 \leq n \leq \tau\,, && (273)
\end{aligned}$$

since otherwise we can truncate τ to the first time instant when inequality (273) is violated. Such truncation can only reduce $\mathbb{E}[\tau]$ and $\mathbb{P}[\hat{W} \neq W]$. It is easy to see that R_n is a P^j-supermartingale (for any j) according to (96) and the classical characterization of capacity

$$C = \max_x D(P_{Y|X=x}||P_Y^*)\,. \tag{274}$$

Consider regular branches $\tilde{P}^j[\cdot]$ and $\tilde{Q}^j[\cdot]$ of conditional probabilities $P^j[\cdot|\mathscr{G}_{\tau_1}]$ and $Q^j[\cdot|\mathscr{G}_{\tau_1}]$. Then, one easily shows that the relative entropy between $\tilde{P}^j$ and $\tilde{Q}^j$ on $\mathscr{G}_\tau$ satisfies

$$D_{\mathscr{G}_\tau}(\tilde{P}^j||\tilde{Q}^j) = \mathbb{E}^j[S_\tau - S_{\tau_1}|\mathscr{G}_{\tau_1}]\,, \tag{275}$$

where here and below $\mathbb{E}^j[\cdot]$ denotes the expectation with respect to P^j. Since R_n is a supermartingale we have further

$$D_{\mathscr{G}_\tau}(\tilde{P}^j||\tilde{Q}^j) \leq C_1 \mathbb{E}^j[\tau - \tau_1|\mathscr{G}_{\tau_1}]\,. \tag{276}$$

Consider now the following chain:

$$\begin{aligned}
d\left(1-\eta_{\tau_1}\left|\right|\eta_{\tau_1}\frac{\pi^{\max}_{\tau_1}}{1-\pi^{\max}_{\tau_1}}\right) &\leq d\left(1-\eta_{\tau_1}\left|\right|\sum_{j=1}^M \pi_{\tau_1}(j)Q^j[\hat{W}=j|\mathscr{G}_{\tau_1}]\right) && (277)\\
&\leq \sum_{j=1}^M \pi_{\tau_1}(j) d(P^j[\hat{W}=j|\mathscr{G}_{\tau_1}]\,||\,Q^j[\hat{W}=j|\mathscr{G}_{\tau_1}]) && (278)\\
&\leq \sum_{j=1}^M \pi_{\tau_1}(j) D_{\mathscr{G}_\tau}(\tilde{P}^j||\tilde{Q}^j) && (279)\\
&\leq C_1 \sum_{j=1}^M \pi_{\tau_1}(j)\mathbb{E}^j[\tau-\tau_1|\mathscr{G}_{\tau_1}] && (280)\\
&= C_1 \mathbb{E}\left[\tau-\tau_1|\mathscr{G}_{\tau_1}\right], && (281)
\end{aligned}$$

where (277) is by (264) applied with $B=\{\hat{W}=j\}$, inequality

$$\begin{aligned}
\sum_{j=1}^M \pi_{\tau_1}(j)\frac{\mathbb{P}[\hat{W}=j, W\neq j|\mathscr{G}_{\tau_1}]}{\mathbb{P}[W\neq j|\mathscr{G}_{\tau_1}]} &\leq \frac{\pi^{\max}_{\tau_1}}{1-\pi^{\max}_{\tau_1}}\sum_{j=1}^M \mathbb{P}[\hat{W}=j, W\neq j|\mathscr{G}_{\tau_1}] && (282)\\
&= \frac{\pi^{\max}_{\tau_1}}{1-\pi^{\max}_{\tau_1}}\eta_{\tau_1}\,, && (283)
\end{aligned}$$

and the fact that the second argument of $d(\cdot||\cdot)$ in the left-hand side of (277) is not larger than the first (according to (273)); (278) is by Jensen's inequality applied to $d(\cdot||\cdot)$ and by an obvious identity

$$1-\eta_{\tau_1} = \sum_{j=1}^M \pi_{\tau_1}(j) P^j[\hat{W}=j|\mathscr{G}_{\tau_1}]\,, \tag{284}$$

(279) is the data-processing for relative entropy, (280) is by (276) and (281) follows since

$$\sum_{j=1}^M \pi_{\tau_1}(j)\mathbb{E}^j[\cdot|\mathscr{G}_{\tau_1}] = \mathbb{E}\left[\cdot|\mathscr{G}_{\tau_1}\right]. \tag{285}$$

By an elementary lower bound on $d(\cdot||\cdot)$ applied to the left-hand side of (277) we obtain from (281)

$$(1-\eta_{\tau_1})\log\frac{1-\pi^{\max}_{\tau_1}}{\eta_{\tau_1}\pi^{\max}_{\tau_1}} - h(\eta_{\tau_1}) \leq C_1\mathbb{E}\left[\tau-\tau_1|\mathscr{G}_{\tau_1}\right]. \tag{286}$$

To estimate the expectation of τ_1 consider another chain

$$\begin{aligned}
\mathbb{E}\left[\log M - F_M(1-\pi_{\tau_1}^{\max})\right] &= \mathbb{E}\left[\log M - F_M(\mathbb{P}[\hat{W}_{\tau_1} \neq W|\mathscr{G}_{\tau_1}])\right] && (287)\\
&= \mathbb{E}\left[d(\mathbb{P}[\hat{W}_{\tau_1} = W|\mathscr{G}_{\tau_1}]||\tfrac{1}{M})\right. && (288)\\
&= \mathbb{E}\left[d(\mathbb{P}[\hat{W}_{\tau_1} = W|\mathscr{G}_{\tau_1}]||\mathbb{Q}[\hat{W}_{\tau_1} = W|\mathscr{G}_{\tau_1}])\right. && (289)\\
&\leq D_{\mathscr{G}_{\tau_1}\vee\sigma\{W\}}(\mathbb{P}||\mathbb{Q}) && (290)\\
&= \mathbb{E}\left[S_{\tau_1}\right] && (291)\\
&\leq C\mathbb{E}\left[\tau_1\right], && (292)
\end{aligned}$$

where (289) is because under $\mathbb{Q}$ W is equiprobable and independent of $\mathscr{G}_{\tau_1}$, (290) is a data-processing inequality applied to measures $\mathbb{P}$ and $\mathbb{Q}$ on the σ-algebra $\mathscr{G}_{\tau_1} \vee \sigma\{W\}$, and (292) is because R_n is a supermartingale.

We now choose

$$\tau_1 = \min\{\tau, \inf\{n \geq 0 : \pi_n^{\max} \geq 1-\xi\}\}. \tag{293}$$

Similar to [14, Proposition 2] one shows that for all n and j we have

$$\lambda_1 \frac{\pi_n(j)}{1-\pi_n(j)} \leq \frac{\pi_{n+1}(j)}{1-\pi_{n+1}(j)} \leq \frac{1}{\lambda_1}\frac{\pi_n(j)}{1-\pi_n(j)}. \tag{294}$$

Since $\lambda_1 < 1$ we can see that regardless of whether $\pi_n^{\max}$ hits level $1-\xi$ before τ or not, we have

$$\frac{1-\pi_{\tau_1}^{\max}}{\pi_{\tau_1}^{\max}} \geq \lambda_1 \frac{\xi}{1-\xi}. \tag{295}$$

On one hand, we have the following estimate

$$\begin{aligned}
C_1\mathbb{E}\left[\tau - \tau_1\right] &\geq \left|\mathbb{E}\left[(1-\eta_{\tau_1})\log\frac{1-\pi_{\tau_1}^{\max}}{\eta_{\tau_1}\pi_{\tau_1}^{\max}} - h(\eta_{\tau_1})\right]\right|^+ && (296)\\
&\geq \left|\mathbb{E}\left[(1-\eta_{\tau_1})\log\frac{\lambda_1\xi}{\eta_{\tau_1}(1-\xi)} - h(\eta_{\tau_1})\right]\right|^+ && (297)\\
&\geq \left|(1-\epsilon)\log\frac{\lambda_1\xi}{\epsilon(1-\xi)} - h(\epsilon)\right|^+, && (298)\\
& && (299)
\end{aligned}$$

where (296) is by (286), (297) is by (295), and (298) is by Jensen's inequality, convexity of $(1-x)\log\frac{1}{x}$ and the trivial identity $\mathbb{P}[\hat{W} \neq W] = \mathbb{E}\left[\eta_\tau\right]$. On the other hand, if we denote an event

$$A = \{\exists\ 0 \leq n \leq \tau : \pi_n^{\max} \geq 1-\xi\}, \tag{300}$$

then

$$\begin{aligned} C\mathbb{E}\left[\tau_1\right] &\geq \log M - \mathbb{E}\left[F_M(1-\pi_{\tau_1}^{\max}\right] && (301)\\ &\geq \log M - \mathbb{P}[A]F_M(\xi) - \mathbb{P}[A^c]F_M(\mathbb{P}[\hat{W}\neq W|A^c]) && (302)\\ &\geq \log M - F_M(\xi) - \min\left\{F_M(\epsilon), \frac{\epsilon}{\xi}\log M\right\}, && (303)\end{aligned}$$

where (301) is by (292); (302) is by concavity of $F_M(x)$ and since on A: $\pi_{\tau_1}^{\max} \geq 1-\xi$, while on A^c: $\pi_{\tau_1}^{\max} = \eta_{\tau_1}$; and (303) is by the bound

$$\mathbb{P}[A^c]F_M(\mathbb{P}[\hat{W}\neq W|A^c]) \leq F_M(\mathbb{P}[\hat{W}\neq W])\,, \tag{304}$$

which follows from concavity of $F_M(\cdot)$ and

$$\begin{aligned} \mathbb{P}[A^c]F_M(\mathbb{P}[\hat{W}\neq W|A^c]) &\leq \mathbb{P}[1-\pi_{\tau}^{\max} > \xi]\log M && (305)\\ &\leq \frac{\epsilon}{\xi}\log M\,, && (306)\end{aligned}$$

which follows by Chebyshev's inequality and (272). Summing (298) and (303) we obtain (98).

■

REFERENCES

[1] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, Vol. 2, No. 3, pp. 8-19, Sept. 1956.

[2] R. L. Dobrushin, "Asymptotic bounds on error probability for transmission over DMC with symmetric transition probabilities," *Theory of Probab. Applicat.*, vol. 7, pp. 283-311, 1962.

[3] M. V. Burnashev, "Data transmission over a discrete channel with feedback. Random transmission time," *Problems of Information Transmission*, vol.12, no.4, pp. 10-30, 1976.

[4] M. V. Burnashev, "Sequential discrimination of hypotheses with control of observations," *Math. USSR, Izvestia*, vol. 15, no. 3, pp. 419-440, 1980.

[5] H. Yamamoto and K. Itoh, "Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback," *IEEE Trans. Inform. Theory*, vol. 25, no. 6, pp. 729-733, Nov. 1979.

[6] N. Shulman, "Communication over an unknown channel via common broadcasting," Ph.D. dissertation, Tel-Aviv Univ., Tel-Aviv, Israel, 2003.

[7] S. C. Draper, B. J. Frey, and F. R. Kschischang, "Efficient variable length channel coding for unknown DMCs," *Proc. 2004 IEEE Int. Symp. Information Theory (ISIT)*, Chicago, IL, USA, June 2004.

[8] A. Tchamkerten and E. Telatar, "A feedback strategy for binary symmetric channels," *Proc. 2002 IEEE Int. Symp. Information Theory (ISIT)*, Lausanne, Switzerland, July 2002.

[9] A. Tchamkerten and E. Telatar, "Optimal feedback schemes over unknown channels," *Proc. 2004 IEEE Int. Symp. Information Theory (ISIT)*, Chicago, IL, USA, June 2004.

[10] A. Tchamkerten and E. Telatar, "Variable length coding over an unknown channel," *IEEE Trans. Inform. Theory*, vol. 52, no. 5, pp. 2126-2145, May 2006.

[11] S. C. Draper, F. R. Kschischang, and B. Frey, "Rateless coding for arbitrary channel mixtures with decoder channel state information," *IEEE Trans. Inform. Theory*, vol. 55, no. 9, pp. 4119-4133, Sep. 2009.

[12] Y. Polyanskiy, H. V. Poor and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2307-2359, May 2010.

[13] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, Jul./Oct. 1948.

[14] P. Berlin, B. Nakiboğlu, B. Rimoldi, and E. Telatar, "A simple converse of Burnashev's reliability function," *IEEE Trans. Inform. Theory*, vol. 55, no. 7, pp. 3074-3080, Jul. 2009.

[15] J. P. M. Schalkwijk and M. Barron, "Sequential signaling under a peak power constraint," *IEEE Trans. Inform. Theory*, vol. 17, no. 5, pp. 278-282, May. 1971.

[16] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 2010.

[17] D. Williams, *Probability with martingales.* Cambridge, UK: Cambridge University Press, 1991.

[18] Y. Polyanskiy, H. V. Poor and S. Verdú, "Dispersion of the Gilbert-Elliot channel," *IEEE Trans. Inform. Theory*, vol. 57, no. 4, pp. 1829-1848, Apr. 2011.

[19] S. Verdú and S. Shamai, "Variable-rate channel capacity," *IEEE Trans. Inform. Theory*, vol. 56, no. 6, pp. 2651-2667, Jun. 2010.

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

[21] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1147-1157, 1994.

[22] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Trans. Inform. Theory*, vol. 54, no. 3, pp. 964-975, Mar. 2008.

[23] A. Sahai, "Why do block length and delay behave differently if feedback is present?" *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 1860 - 1886, May 2008.

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

Yury Polyanskiy (S'08-M'10) received the M.S. degree (with honors) in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia in 2005 and a Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ in 2010.

In 2000-2005, he was with the Department of Surface Oilfield Equipment, Borets Company LLC, where he rose to the position of Chief Software Designer. His research interests include information theory, coding theory and the theory of random processes.

Dr. Polyanskiy won a silver medal at the 30th International Physics Olympiad (IPhO), held in Padova, Italy. He was a recipient of the Best Student Paper Awards at the 2008 and 2010 IEEE International Symposiums on Information Theory (ISIT).

H. Vincent Poor (S'72-M'77-SM'82-F'87) received the Ph.D. degree in electrical engineering and computer science from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is the Dean of Engineering and Applied Science, and the Michael Henry Strater University Professor of Electrical Engineering. Dr. Poor's research interests are in the areas of stochastic analysis, statistical signal processing and information theory, and their applications in wireless networks and related fields. Among his publications in these areas are *Quickest Detection* (Cambridge University Press, 2009), co-authored with Olympia Hadjiliadis, and *Information Theoretic Security* (Now Publishers, 2009), co-authored with Yingbin Liang and Shlomo Shamai.

Dr. Poor is a member of the National Academy of Engineering, a Fellow of the American Academy of Arts and Sciences, and an International Fellow of the Royal Academy of Engineering (U.K.). He is also a Fellow of the Institute of Mathematical Statistics, the Optical Society of America, and other organizations. In 1990, he served as President of the IEEE Information Theory Society, in 2004-07 as the Editor-in-Chief of these *Transactions*, and in 2009 as General Co-chair of the IEEE International Symposium on Information Theory, held in Seoul, South Korea. He received a Guggenheim Fellowship in 2002 and the IEEE Education Medal in 2005. Recent recognition of his work includes, the 2009 Edwin Howard Armstrong Achievement Award of the IEEE Communications Society, the 2010 IET Ambrose Fleming Medal for Achievement in Communications, the 2011 IEEE Eric E. Sumner Award, and an honorary D. Sc. from the University of Edinburgh, awarded in 2011.

Sergio Verdú (S'80-M'84-SM'88-F'93) is the Eugene Higgins Professor of Electrical Engineering at Princeton University.

A member of the National Academy of Engineering, Verdú is the recipient of the 2007 Claude Shannon Award and the 2008 IEEE Richard Hamming Medal. He was awarded a Doctorate Honoris Causa from the Universitat Politècnica de Catalunya in 2005.

His research has received several awards including the 1998 Information Theory Outstanding Paper Award, the Information Theory Golden Jubilee Paper Award, and the 2006 Joint Communications/Information Theory Paper Award.

Sergio Verdú is currently Editor-in-Chief of *Foundations and Trends in Communications and Information Theory.*