Upper Bounds on the Feedback Error Exponent of Channels With States and With Memory

Mohsen Heidari
CS Department
Purdue University
Email: mheidari@purdue.edu

Achilleas Anastasopoulos
ECE Department
University of Michigan
Email: anastas@umich.edu

S. Sandeep Pradhan
ECE Department
University of Michigan
Email: pradhanv@umich.edu

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834603

***Abstract*—As a class of state-dependent channels, Markov channels have been long studied in information theory for characterizing the feedback capacity and error exponent. This paper studies a more general variant of such channels where the state evolves via a general stochastic process, not necessarily Markov or ergodic. The states are assumed to be unknown to the transmitter and the receiver, but the underlying probability distributions are known. For this setup, we derive an upper bound on the feedback error exponent and the feedback capacity with variable length codes (VLCs). The bounds are expressed in terms of the directed mutual information and directed relative entropy. The bounds on the error exponent reduce to Burnashev's expression for discrete memoryless channels. Our method relies on tools from the theory of martingales to analyze a stochastic process defined based on the entropy of the message given the past channel's outputs.**

I. Introduction

Communications over channels with feedback has been a longstanding problem in information theory literature. The early works on discrete memoryless channels (DMCs) pointed to a negative answer as to whether feedback can increase the capacity [1]. Feedback, though, improves the channel's error exponent — the maximum attainable exponential rate of decay of the error probability. The improvements are obtained using variable length codes (VLCs), where the communication length depends on the channel's realizations. In a seminal work, Burnashev [2] completely characterized the error exponent of DMCs with noiseless and casual feedback. This characterization has a simple, yet intuitive, form:

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

where R is the (average) rate of transmission, C is the channel's capacity, and C_1 is the maximum exponent for binary hypothesis testing over the channel. It is equal to the maximal relative entropy between conditional output distributions. The Burnashev's exponent can significantly exceed the sphere-packing exponent for no-feedback communications as it approaches capacity with a nonzero slope. The use of VLCs is shown to be essential to establish these results, as no improvements are gained using fixed-length codes [3]–[5].

This result led to the question as to whether the feedback improves the capacity or error exponent of more general channels, modeling non-traditional communications involving memory and intersymbol interference (ISI). Among such models are channels with states where the transition probability of the channel varies depending on its state which itself evolves based on the past inputs and channel's realizations. Depending on the variants of this formulation, the agents may have no knowledge about the state (e.g., arbitrarily varying channels), or they may exactly know the state [6]. Feedback can improve the error exponent when the state is known at the transmitter and the receiver. Notably, Como, *et al.* [7] extended Burnashev-type exponent to finite-state ergodic Markov channels with known state and derived a similar form as in (1). The error exponent for channels with more general state evolution is still unknown; only the feedback capacity, when restricted to fixed-length codes, is known [8].

This paper studies the feedback error exponent for channels with more general state evolution and allowing VLCs. More precisely, we study discrete channels with states where the state evolves via an arbitrary stochastic process (not necessarily ergodic or Markov) depending on the past realizations. Furthermore, the realization of the states is assumed to be unknown, but the transmitter or the receiver may know the underlying probability distribution governing the evolution of the state. However, noiseless feedback is available at the transmitter with one unit of delay. The main contributions are twofold. First, we prove an upper bound on the error exponent of such channels, which has the familiar form

$$E(R) \leq \sup_{N>0} \sup_{P^N \in \mathcal{P}^N} D(P^N)(1 - \frac{R}{I(P^N)}),$$

where D is the directed relative entropy, I is the directed mutual information, and $\mathcal{P}^N$ is a collection of "feasible" probability distributions. As a special case, the bound reduces to Burnashev's expression when the channel is DMC. Second, we introduce an upper bound on the feedback capacity of VLCs for communications over these channels with stochastic states. This upper bound generalizes the results of Tatikonda and Mitter [8] and Permuter *et al.* [9] where fixed-length codes are studied. Our approach relies on the analysis of the entropy of the stochastic process defined based on the entropy of the message given the past channel's output. We analyze the drift of the entropy via tools from the theory of martingales.

This work was supported by NSF grant CCF-2132815.

Related works on the capacity and error exponent of channels with feedback are extensive. Yamamoto and Itoh [10] introduced a two-phase iterative for achieving the Burnashev exponent. Berlin *et al.* [11] provided a simpler proof of Burnashev's expression for DMCs. The error exponent of DMCs with feedback and cost constraints is studied in [12]. Channels with state and feedback have been studied under various frameworks on the evolution model of the states and whether they are known at the transmitter or the receiver. On one extreme of such models are arbitrarily varying channels [13]. The feedback capacity of these channels for fixed-length codes is derived in [8]. Tchamkerten and Telatar [14] studied the universality of Burnashev error exponent. They considered communication setups where the parties have no exact knowledge of the channel's statistics but know it belongs to a certain class of DMCs. The authors proved that no zero-rate coding scheme achieves the Burnashev's exponent simultaneously for all the DMC's in the class. However, they showed positive results for two families of such channels (e.g., binary symmetric and Z) [15]. Another class of channels with states are Markov channels that have been studied extensively for deriving their capacity [6], [16], [17], and error exponent using fixed-length codes [8]. A lower bound on the error exponent of unifilar channels is derived [18], where the state is a deterministic function of the previous ones. Other variants of this problem have been studied, including continuous-alphabet channels [19], [20], and multi-user channels [21], [22].

II. Problem Formulation and Definitions

The formal definitions are presented in this section. For shorthand, we use $[1:M]$ to denote $\{1, 2, ..., M\}$.

A discrete channel with stochastic state has three finite sets $\mathcal{X}, \mathcal{Y}$, and $\mathcal{S}$ representing the input, output, and state of the channel, respectively. Consider a collection of channels $\mathcal{Q} := \{Q(\cdot|\cdot, s) : s \in \mathcal{S}\}$, indexed by $s \in \mathcal{S}$, where each element $Q(\cdot|\cdot, s) : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ is the transition probability of the channel at state s. The states $\{S_t\}_{t>0}$, evolve according to $P_{S_t|S^{t-1},X^{t-1}}, t > 0$ depending on the past inputs and state realizations. As a result, after t uses of the channel with x^{t-1}, s^{t-1}, y^{t-1} being the channels input, state and output, the next output is given by

$$P(s_t, y_t|x^{t-1}, s^{t-1}, y^{t-1}) = P_{t,S}(s_t|s^{t-1}, x^{t-1})Q(y_t|x_t, s_t).$$

Such evolution of the states induces memory over time as it depends on past inputs.

After each use of the channel, the output of the channel y_t as feedback is available at the transmitter with one unit of delay. Moreover, we allow VLCs for communications where the transmitter nor the receiver do not know the state of the channel. More precisely, the setup is defined as follows.

Definition 1. *An (M, N)-VLC for communications over a channel $\mathcal{Q}$ with states and feedback is defined by*

- *A message W with uniform distribution over $[1 : M]$.*
- *Encoding functions $e_t : [1 : M] \times \mathcal{Y}^{t-1} \to \mathcal{X}, t \in \mathbb{N}$.*
- *Decoding functions $d_t : \mathcal{Y}^t \to [1 : M], t \in \mathbb{N}$.*
- *A stopping time T with respect to (w.r.t) the filtration $\mathcal{F}_t$ defined as the σ-algebra of Y^t for $t \in \mathbb{N}$. Furthermore, it is assumed that T is almost surely bounded as $T \leq N$.*

For technical reasons, we study a class of (M, N)-VLCs for which the parameter N grows sub-exponentially with $\log M$, that is $N \leq (\log M)^m$ for some fixed number m. An example is the sequence $(M^{(n)}, N^{(n)})$-VLCs, $n \geq 1$, where $M^{(n)} = 2^{nr_1}, N^{(n)} \leq n^m$, with $r_1, r_2, m > 0$ being fixed parameters.

In what follows, for any (M, N)-VLC, we define average rate, error probability, and error exponent. Given a message W, the tth output of the transmitter is denoted by $X_t = e_t(W, Y^{t-1})$, where Y^{t-1} is the noiseless feedback up to time t. Let $\hat{W}_t = d_t(Y^t)$ represent the estimate of the decoder about the message. Then, at the end of the stopping time T, the decoder declares $\hat{W}_T$ as the decoded message. The average rate and (average) probability of error for a VLC are defined as

$$R \triangleq \frac{\log_2 M}{\mathbb{E}[T]}, \quad P_e \triangleq \mathbb{P}\Big\{\hat{W}_T \neq W\Big\}.$$

Definition 2. *A rate R is achievable for a given channel with stochastic states, if there exists a sequence of $(M^{(n)}, N^{(n)})$-VLCs such that*

$$\limsup_{n\to\infty} P_e^{(n)} = 0, \qquad \limsup_{n\to\infty} \frac{\log M^{(n)}}{\mathbb{E}[T^{(n)}]} \geq R,$$

and $N^{(n)} \leq (n)^m, \forall n > 1$, where m is fixed. The feedback capacity, $\mathcal{C}_F^{VLC}$, is the convex closure of all achievable rates.

Naturally, the error exponent of a VLC with probability of error P_e and stopping time T is defined as $E \triangleq -\frac{\log_2 P_e}{\mathbb{E}[T]}$. The following definition formalizes this notion.

Definition 3. *An error exponent function $E(R)$ is said to be achievable for a given channel, if for any rate $R > 0$ there exists a sequence of $(M^{(n)}, N^{(n)})$-VLCs such that*

$$\liminf_{n\to\infty} -\frac{\log P_e^{(n)}}{\mathbb{E}[T^{(n)}]} \geq E(R), \qquad \limsup_{n\to\infty} \frac{\log M^{(n)}}{\mathbb{E}[T^{(n)}]} \geq R,$$

and $\limsup_{n\to\infty} M^{(n)} = \infty$ with $N^{(n)} \leq (n)^m, \forall n > 1$, where m is fixed. The reliability function is the supremum of all achievable reliability functions $E(R)$.

III. Main Results

We start with deriving an upper bound on the feedback capacity of channels with stochastic states and allowing VLCs. The expressions are based on the *directed information* as introduced in [23] and defined as

$$I(X^n \to Y^n) \triangleq \sum_{i=1}^{n} I(X_i; Y_i|Y^{i-1}). \tag{2}$$

We further extend this notion to variable-length sequences. Consider a stochastic process $\{(X_t, Y_t)\}_{t>0}$ and let T be a (bounded) stopping time w.r.t an induced filtration $\mathcal{F}_t, t > 0$. Then, the directed mutual information is defined as

$$I(X^T \to Y^T) \triangleq \mathbb{E}\big[\sum_{t=1}^{T} I(X_t; Y_t \mid \mathcal{F}_{t-1})\big]. \tag{3}$$

Now, we are ready for an upper bound on the feedback capacity. For any integer N, let $\mathcal{P}^N$ be the set of all N-letter distributions P_{X^N,S^N,Y^N} on $\mathcal{X}^N \times \mathcal{S}^N \times \mathcal{Y}^N$ that factors as

$$\prod_{\ell=1}^{N} P_\ell(x_\ell|x^{\ell-1}, y^{\ell-1}) P_\ell(s_\ell|s^{\ell-1}, x^{\ell-1}) Q(y_\ell|x_\ell, s_\ell). \quad (4)$$

Theorem 1. *The feedback capacity of a channel with stochastic states is bounded as*

$$\mathcal{C}_F^{VLC} \leq \sup_{N>0} \sup_{P^N \in \mathcal{P}^N} \sup_{\textit{stop time } T: T \leq N} \frac{1}{\mathbb{E}[T]} I(X^T \to Y^T).$$

Observe that for a trivial stopping time $T = N$, the bound reduces to that for fixed-length codes as given in [8].

A. Upper Bound on the Error Exponent

We need a notation to proceed. Consider a pair of random sequences $(X^n, Y^n) \sim P_{X^n Y^n}$. Let X_r^* be the MAP estimation of X_r from observation Y^{r-1}, that is $X_r^* = \arg\max_x \mathbb{P}\left\{X_r = x | Y^{r-1} = y^{r-1}\right\}$. Also, let $\bar{Q}_r = P_{Y_r|X_r, Y^{r-1}}$ which is the effective channel (averaged over possible states) from the transmitter's perspective at time r. With this notation, we define the directed KL-divergence as

$$D(X^n \to Y^n) \triangleq \max_{x^n} \sum_{r=1}^{n} D_{KL}\Big(\bar{Q}_r(\cdot|X_r^*, Y^{r-1}) \Big\| \bar{Q}_r(\cdot|x_r, Y^{r-1}) \;\Big|\; Y^{r-1}\Big).$$

Intuitively, $D(X^n \to Y^n)$ measures the sum of the expected "distance" between the channel's probability distribution conditioned on the MAP symbol versus the worst symbol across different times $r \in [1:n]$.

Theorem 2. *The error exponent of a channel with stochastic states is bounded as*

$$E(R) \leq \sup_{N \in \mathbb{N}} \sup_{P^N \in \mathcal{P}^N} \sup_{T: T \leq N} \sup_{T_1: T_1 \leq T} D(P^N)\Big(1 - \frac{R}{I(P^N)}\Big),$$

where T, T_1 are stopping times, and

$$I(P^N) = \frac{1}{\mathbb{E}[T_1]} I(X^{T_1} \to Y^{T_1}),$$
$$D(P^N) = \frac{1}{\mathbb{E}[T - T_1]} D(X_{T_1+1}^T \to Y_{T_1+1}^T).$$

In the next section, we present our proof techniques.

IV. Proof of Theorem 2

The proof follows by a careful study of the drift of the entropy of the message W conditioned on the channel's output at each time t. Define the following random process:

$$H_t = H(W|\mathcal{F}_t), t > 0, \quad (5)$$

where $\mathcal{F}_t$ is the σ-algebra of Y^t. We show that H_t drifts in three phases: (i) linear drift (data phase) until reaching a small value (ϵ); (ii) fluctuation phase with values around ϵ; and (iii) logarithmic drift (hypothesis testing phase) till the end. We derive bounds on the expected slop of the drifts and

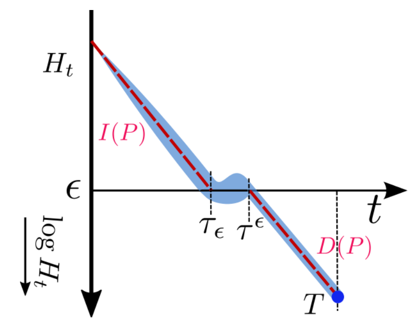

Fig. 1. Entropy drifts over t. The first part till τ_ϵ, as in (6), is the linear drift with the expected slop $I(P)$ (dashed line). From τ_ϵ to τ^ϵ, as in (7), are the fluctuations around ϵ (shaded region). Then, from τ^ϵ to T is the logarithmic drift ($\log H_t$) with the expected slop of $D(P)$ (the second dashed line).

prove that the length of the fluctuation phase is asymptotically negligible as compared to the communication length (Fig. 1).

More precisely, we have the following argument by defining a pruned time random process $\{t_n\}_{n>0}$. First, for any $\epsilon \geq 0$ and $N \in \mathbb{N}$ define the following random variables

$$\tau_\epsilon \triangleq \inf\{t > 0 : H_t \leq \epsilon\} \wedge N \quad (6)$$
$$\tau^\epsilon \triangleq \sup\{t > 0 : H_{t-1} \geq \epsilon\} \wedge N \quad (7)$$

Then, the pruned time process is defined as

$$t_n \triangleq \begin{cases} n & \text{if } n < \tau_\epsilon \\ n \vee \tau^\epsilon & \text{if } \tau_\epsilon \leq n \leq N \\ N & \text{if } n > N \end{cases} \quad (8)$$

Note that τ_ϵ is a stopping time with respect to $\{H_t\}_{t>0}$, but this is not the case for τ^ϵ.

Lemma 1. *Suppose a non-negative random process $\{H_r\}_{r>0}$ has the following properties w.r.t a filtration $\mathcal{F}_r, r > 0$,*

$$\mathbb{E}[H_{r+1} - H_r|\mathcal{F}_r] \geq -k_{1,r+1}, \quad \text{if } H_r \geq \epsilon, \quad (9a)$$
$$\mathbb{E}[\log H_{r+1} - \log H_r|\mathcal{F}_r] \geq -k_{2,r+1} \quad \text{if } H_r < \epsilon \quad (9b)$$
$$|\log H_{r+1} - \log H_r| \leq k_3 \quad (9c)$$
$$|H_{r+1} - H_r| \leq k_4 \quad (9d)$$

where $k_{1,r}, k_{2,r}, k_3, k_4$ are non-negative numbers and $k_{1,r} \leq k_{2,r}$ for all $r > 0$. Given $\epsilon \in (0,1)$, and $I \geq D > 0$, let

$$Z_t \triangleq \frac{H_t - \epsilon}{I} \mathbb{1}\{H_t \geq \epsilon\} + \Big(\frac{\log \frac{H_t}{\epsilon}}{D} + f(\log \frac{H_t}{\epsilon})\Big) \mathbb{1}\{H_t < \epsilon\},$$

where $f(y) = \frac{1 - e^{\lambda y}}{\lambda D}$ with $\lambda > 0$. Further define $\{S_t\}_{t>0}$ as

$$S_t \triangleq \sum_{r=1}^{t \wedge \tau_\epsilon} \frac{k_{1,r}}{I} + \sum_{r=t \wedge \tau_\epsilon + 1}^{t \wedge \tau^\epsilon} \frac{k_4}{I} \mathbb{1}\{H_{r-1} \geq \sqrt{\epsilon}\} + \sum_{r=t \wedge \tau^\epsilon + 1}^{t} \frac{k_{2,r}}{D} + \sqrt{\epsilon}\frac{N}{I} \mathbb{1}\{t \geq \tau^\epsilon\}.$$

Let $\{t_n\}_{n>0}$ be as in (8) but w.r.t $\{H_r\}_{r>0}$. Lastly define the random process $\{L_n\}_{n>0}$ as $L_n \triangleq Z_{t_n} + S_{t_n}$. Then, for small enough $\lambda > 0$ the process $\{L_n\}_{n>0}$ is a sub-martingale with respect to the time pruned filtration $\mathcal{F}_{t_n}, n > 0$.

Proof: We consider three cases depending on n.

Case (a). $n < \tau_\epsilon - 1$**:** Because of (8), $t_n = n$ and $t_{n+1} = n+1$. As n did not reach τ_ϵ, then $H_n > \epsilon$ and $H_{n+1} > \epsilon$. Therefore,

$$L_n = Z_{t_n} + S_{t_n} = Z_n + S_n = \frac{H_n - \epsilon}{I} + \sum_{r=1}^{n} \frac{k_{1,r}}{I}$$
$$L_{n+1} = Z_{n+1} + S_{n+1} = \frac{H_{n+1} - \epsilon}{I} + \sum_{r=1}^{n+1} \frac{k_{1,r}}{I}. \tag{10}$$

As a result, the difference between L_n and L_{n+1} equals to

$$\mathbb{E}[(L_{n+1} - L_n)\mathbb{1}\{n < \tau_\epsilon - 1\}|y^{t_n}] = \mathbb{E}[L_{n+1} - L_n|y^n]\mathbb{1}\{n < \tau_\epsilon - 1\},$$

where the equality holds as $t_n = n$, and τ_ϵ is a stopping time implying that $\mathbb{1}\{n < \tau_\epsilon - 1\}$ is a function of y^n. Next, from (10), the difference term above is bounded as

$$\mathbb{E}[L_{n+1} - L_n|y^n] = \mathbb{E}[\frac{H_{n+1} - H_n}{I} + \frac{k_{1,n+1}}{I}|y^n] = \frac{\mathbb{E}[H_{n+1} - H_n|y^n]}{I} + \frac{k_{1,n+1}}{I} \geq 0,$$

where the last inequality follows from (9a). As a result, we proved that $\mathbb{E}[(L_{n+1} - L_n)\mathbb{1}\{n < \tau_\epsilon - 1\}|y^{t_n}] \geq 0$.

Case (b). $n = \tau_\epsilon - 1$**:** In this case, $t_n = n$ implying that $H_n > \epsilon$ and $t_{n+1} = (n+1) \vee \tau^\epsilon$. Furthermore, since, $n+1 = \tau_\epsilon \leq \tau^\epsilon$, then $t_{n+1} = \tau^\epsilon$. Consequently, we obtain that

$$L_n = Z_n + S_n = \frac{H_n - \epsilon}{I} + \sum_{r=1}^{n} \frac{k_{1,r}}{I}$$
$$L_{n+1} = Z_{\tau^\epsilon} + S_{\tau^\epsilon} = (\frac{H_{\tau^\epsilon} - \epsilon}{I})\mathbb{1}\{H_{\tau^\epsilon} \geq \epsilon\} + (\frac{\log H_{\tau^\epsilon} - \log\epsilon}{D} + f(\log\frac{H_{\tau^\epsilon}}{\epsilon}))\mathbb{1}\{H_{\tau^\epsilon} < \epsilon\} + \sum_{r=1}^{\tau_\epsilon} \frac{k_{1,r}}{I} + \sum_{r=\tau_\epsilon+1}^{\tau^\epsilon} \frac{k_4}{I}\mathbb{1}\{H_{r-1} \geq \sqrt{\epsilon}\} + \sqrt{\epsilon}\frac{N}{I}.$$

Note that Z_{τ^ϵ} does not necessarily equal to the logarithmic part. The reason is that τ^ϵ is pruned by N as in (7). Thus, H_{τ^ϵ} can be greater than ϵ when $\tau^\epsilon = N$. We proceed by bounding Z_{τ^ϵ}. Note that, for small enough λ

$$\frac{\epsilon}{I}(e^y - 1) - \frac{y}{D} < f(y), \qquad -k_3 < y < 0. \tag{11}$$

Applying inequality (11) with $y = \log\frac{H_{\tau^\epsilon}}{\epsilon}$, we can write that

$$Z_{\tau^\epsilon} > (\frac{H_{\tau^\epsilon} - \epsilon}{I})\mathbb{1}\{H_{\tau^\epsilon} \geq \epsilon\} + (\frac{H_{\tau^\epsilon} - \epsilon}{I})\mathbb{1}\{H_{\tau^\epsilon} < \epsilon\} = \frac{H_{\tau^\epsilon} - \epsilon}{I} \tag{12}$$

Consequently, the difference $L_{n+1} - L_n$ satisfies the following

$$\mathbb{E}[(L_{n+1} - L_n)\mathbb{1}\{n = \tau_\epsilon - 1\}|y^{t_n}] = \mathbb{E}[L_{n+1} - L_n|y^n]\mathbb{1}\{n = \tau_\epsilon - 1\} \geq \mathbb{E}\Big[\frac{H_{\tau^\epsilon} - H_n}{I} + \frac{k_{1,\tau_\epsilon}}{I} + \sum_{r=\tau_\epsilon+1}^{\tau^\epsilon} \frac{k_4}{I}\mathbb{1}\{H_{r-1} \geq \sqrt{\epsilon}\} + \sqrt{\epsilon}\frac{N}{I}\Big|y^n\Big]\mathbb{1}\{n = \tau_\epsilon - 1\} \tag{13}$$

Next, we bound the first term above as

$$H_{\tau^\epsilon} - H_n = H_{n+1} - H_n + \sum_{r=n+2}^{\tau^\epsilon} (H_r - H_{r-1}),$$

where in the first equality, we add and subtract the intermediate terms $H_r, n+1 \leq r \leq \tau^\epsilon - 1$. Next, we substitute the above terms in the RHS of (13). As $n+2 = \tau_\epsilon + 1$, then

$$(13) = \mathbb{E}\Big[\frac{H_{n+1} - H_n}{I} + \frac{k_{1,\tau_\epsilon}}{I} + \sum_{r=\tau_\epsilon+1}^{\tau^\epsilon} \big(\frac{H_r - H_{r-1}}{I} + \frac{k_4}{I}\mathbb{1}\{H_{r-1} \geq \sqrt{\epsilon}\}\big) + \sqrt{\epsilon}\frac{N}{I}\Big|y^n\Big]\mathbb{1}\{n = \tau_\epsilon - 1\}$$
$$\geq \mathbb{E}\Big[\sum_{r=\tau_\epsilon+1}^{\tau^\epsilon} \big(\frac{H_r - H_{r-1}}{I} + \frac{k_4}{I}\mathbb{1}\{H_{r-1} \geq \sqrt{\epsilon}\}\big) + \sqrt{\epsilon}\frac{N}{I}\Big|y^n\Big]\mathbb{1}\{n = \tau_\epsilon - 1\}, \tag{14}$$

where the inequality holds from (9a) and the fact that $n+1 = \tau_\epsilon$. Next, by factoring I and the indicator function inside the expectation, we have the following chain of inequalities

$$(14) = \frac{1}{I}\mathbb{E}\Big[\sum_{r=\tau_\epsilon+1}^{\tau^\epsilon} \big((H_r - H_{r-1}) + k_4\big)\mathbb{1}\{H_{r-1} \geq \sqrt{\epsilon}\} + \big((H_r - H_{r-1})\mathbb{1}\{H_{r-1} < \sqrt{\epsilon}\}\big) + \sqrt{\epsilon}N\Big|y^n\Big]\mathbb{1}\{n = \tau_\epsilon - 1\}$$
$$\overset{(a)}{\geq} \frac{1}{I}\mathbb{E}\Big[\sum_{r=\tau_\epsilon+1}^{\tau^\epsilon} \big((H_r - H_{r-1})\mathbb{1}\{H_{r-1} < \sqrt{\epsilon}\}\big) + \sqrt{\epsilon}N\Big|y^n\Big]\mathbb{1}\{n = \tau_\epsilon - 1\}$$
$$\overset{(b)}{\geq} \frac{1}{I}\mathbb{E}\Big[\Big(\sum_{r=\tau_\epsilon+1}^{\tau^\epsilon} -H_{r-1}\mathbb{1}\{H_{r-1} < \sqrt{\epsilon}\}\Big) + \sqrt{\epsilon}N\Big|y^n\Big]\mathbb{1}\{n = \tau_\epsilon - 1\} \tag{15}$$

where (a) is due to (9d), and (b) holds as $H_r \geq 0$.

Next, from $H_{r-1}\mathbb{1}\{H_{r-1} < \epsilon\} < \epsilon$, we have that

$$(15) > \frac{1}{I}\mathbb{E}\Big[\big(\sum_{r=\tau_\epsilon+1}^{\tau^\epsilon} -\sqrt{\epsilon}\big) + \sqrt{\epsilon}N\Big|y^n\Big]\mathbb{1}\{n = \tau_\epsilon - 1\}$$
$$\overset{(c)}{\geq} \frac{1}{I}\mathbb{E}\Big[\big(\sum_{r=1}^{N} -\sqrt{\epsilon}\big) + \sqrt{\epsilon}N\Big|y^n\Big]\mathbb{1}\{n = \tau_\epsilon - 1\} \geq 0,$$

where (c) holds as $\tau^\epsilon \leq N$. To sum up, we proved that

$$\mathbb{E}[(L_{n+1} - L_n)\mathbb{1}\{n = \tau_\epsilon - 1\}|y^{t_n}] \geq 0.$$

Case (c). $n \geq \tau_\epsilon$**:** This case follows from a similar argument that is given in the full version of the paper. ■

Now, we show that $\{H_t\}_{t>0}$ as in (5) has the conditions in Lemma 1. First (9a) holds because of the following lemma.

Lemma 2. *Given any* (M, N)*-VLC, the following inequality holds almost surely for* $1 \leq r \leq N$

$$\mathbb{E}[H_r - H_{r-1}|\mathcal{F}_{r-1}] = -J_r, \tag{16}$$

where $J_r \triangleq I(X_r;Y_r|\mathcal{F}_{r-1})$ *with the induced* $P_{X^N,Y^N} \in \mathcal{P}^N$.

Condition (9b) holds as a result of the following lemma.

Lemma 3. *For any* (M,N)*-VLC and* $\epsilon \in [0,\frac{1}{2}]$*, if* $H_r < \epsilon$*, then the following inequality holds almost surely*

$$\mathbb{E}[\log H_r - \log H_{r-1}|\mathcal{F}_{r-1}] \geq -D_r + O(h_b^{-1}(\epsilon)), \tag{17}$$

where and D_r *is a function of* y^{r-1} *and is defined as*

$$D_r \triangleq \max_{x\in\mathcal{X}} D_{KL}\Big(\bar{Q}_r(\cdot|x_r^*, y^{r-1}) \parallel \bar{Q}_r(\cdot|x, y^{r-1})\Big), \tag{18}$$

where $\bar{Q}_r = P_{Y_r|X_r,Y^{r-1}}$ *is the average channel from the transmitter's perspective, and* x_r^* *is the MAP input symbol given by* $x_r^* = \arg\max_x \mathbb{P}\left\{X = x|Y^{r-1} = y^{r-1}\right\}$.

Condition (9c) is a direct consequence of Lemma 4 in [2]:

Remark 1. *If* $Q(\cdot|\cdot,\cdot)$ *are positive everywhere then* $|\log H_r - \log H_{r-1}| \leq \eta$*, where*

$$\eta \triangleq \max_{x_1,x_2\in\mathcal{X}} \max_{s_1,s_2\in\mathcal{S}} \max_{y\in\mathcal{Y}} \log\frac{Q(y|x_1,s_1)}{Q(y|x_2,s_2)}.$$

Lastly, (9d) holds as $H_r \leq \log M$ which implies that

$$|H_r - H_{r-1}| \leq \max\{H_r, H_{r-1}\} \leq \log M.$$

Thus, we apply Lemma 1 on $\{H_t\}_{t>0}$ with

$$k_{1,r} = J_r,\ k_{2,r} = D_r,\ k_3 = \eta,\ k_4 = \log M,$$

and constants I, D to be specified later. Therefore, $\{L_n\}_{n>0}$ as in the lemma is a sub-martingale w.r.t $\mathcal{F}_{t_n}, n > 0$.
Connection to the error exponent: Since $\{L_n\}_{n>0}$ is a sub-martingale, then $L_0 \leq \mathbb{E}[L_{T\vee\tau^\epsilon}]$, where T is the stopping time used in the VLC and τ^ϵ is as in (7). Note that $L_0 = \frac{\log M}{I}$. In what follows, we analyze $\mathbb{E}[L_{T\vee\tau^\epsilon}]$.

By definition $L_n = Z_{t_n} + S_{t_n}$. Since, $T \leq N$, then from (8) we have that $t_{T\vee\tau^\epsilon} = (T\vee\tau^\epsilon)\vee\tau^\epsilon = T\vee\tau^\epsilon$. Therefore,

$$\begin{aligned}
\mathbb{E}[L_{T\vee\tau^\epsilon}] &= \mathbb{E}\Big[\frac{H_{T\vee\tau^\epsilon}-\epsilon}{I}\mathbb{1}\{H_{T\vee\tau^\epsilon} \geq \epsilon\} + \Big(\frac{\log(H_{T\vee\tau^\epsilon}/\epsilon)}{D} \\
&\qquad + f\big(\log\frac{H_{T\vee\tau^\epsilon}}{\epsilon}\big)\Big)\mathbb{1}\{H_{T\vee\tau^\epsilon} < \epsilon\} + S_{T\vee\tau^\epsilon}\Big] \\
&\overset{(a)}{\leq} \mathbb{E}\Big[\frac{H_{T\vee\tau^\epsilon}+\epsilon}{I} + \frac{\log(H_{T\vee\tau^\epsilon}/\epsilon)}{D} + f\big(\log\frac{H_{T\vee\tau^\epsilon}}{\epsilon}\big)\Big] \\
&\qquad + \mathbb{E}\big[S_{T\vee\tau^\epsilon}\big] \\
&\overset{(b)}{\leq} \frac{\mathbb{E}\big[H_{T\vee\tau^\epsilon}\big]+\epsilon}{I} + \frac{\log(\mathbb{E}[H_{T\vee\tau^\epsilon}]/\epsilon)}{D} + \frac{1}{\lambda D} + \mathbb{E}\big[S_{T\vee\tau^\epsilon}\big]
\end{aligned}$$

where (a) follows by changing $-\epsilon$ to $+\epsilon$ for the linear part and from the following inequality for the logarithmic part

$$(\log x - \log\epsilon)\mathbb{1}\{x < \epsilon\} \leq \log x - \log\epsilon.$$

Inequality (b) follows from Jensen's inequality, concavity of $\log(x)$ and the inequality $f(y) \leq \frac{1}{\lambda D}$. Next, we bound $\mathbb{E}\big[H_{T\vee\tau^\epsilon}\big]$. As conditioning reduces the entropy, then

$$H_{T\vee\tau^\epsilon} = H(W|Y^{T\vee\tau^\epsilon}) \leq H(W|Y^T) = H_T,$$

where the inequality holds as Y^T is a function of $Y^{T\vee\tau^\epsilon}$. Next, Fano's inequality implies that

$$\mathbb{E}\big[H_{T\vee\tau^\epsilon}\big] \leq \mathbb{E}\big[H_T\big] = \mathbb{E}[H(W|Y^T)] \leq \alpha(P_e), \tag{19}$$

where $\alpha(P_e) = h_b(P_e) + P_e\log(M)$ is the Fano's expression. Therefore, from (19), we obtain that

$$\frac{\log M}{I} \leq \frac{\alpha(P_e)+\epsilon}{I} + \frac{\log\alpha(P_e)-\log\epsilon}{D} + \frac{1}{\lambda D} + \mathbb{E}\big[S_{T\vee\tau^\epsilon}\big].$$

Therefore, rearranging the terms and multiplying by D and dividing by $\mathbb{E}[T]$ give the following

$$\frac{-\log\alpha(P_e)}{\mathbb{E}[T]} \leq D\Big(\frac{\mathbb{E}\big[S_{T\vee\tau^\epsilon}\big]}{\mathbb{E}[T]} - \frac{R}{I}\Big) + U(P_e, M, \epsilon), \tag{20}$$

where we used the fact that $\frac{\log M}{\mathbb{E}[T]} \geq R$, and that

$$U(P_e, M, \epsilon) = R\big(\frac{\alpha(P_e)+\epsilon}{I\log M} + \frac{-\log\epsilon + 1/\lambda}{D\log M}\big). \tag{21}$$

Next, it is not difficult to show that $-\log\alpha(P_e) \geq (-\log P_e)(1-\Delta)$, where

$$\Delta = \frac{\log\big(-\log P_e + 2 + \log M\big)}{-\log P_e}. \tag{22}$$

Therefore, we get the following bound on the error exponent

$$\frac{-\log P_e}{\mathbb{E}[T]} \leq \frac{D}{1-\Delta}\Big(\frac{\mathbb{E}\big[S_{T\vee\tau^\epsilon}\big]}{\mathbb{E}[T]} - \frac{R}{I} + U(P_e, M, \epsilon)\Big). \tag{23}$$

Next, we find appropriate I and D so that $\mathbb{E}\big[S_{T\vee\tau^\epsilon}\big] \approx \mathbb{E}[T]$. Further, we show that Δ and $U(P_e, M, \epsilon)$ converge to zero for any sequence of VLCs satisfying Definition 3.

Lemma 4. *Given* $\epsilon > \alpha(P_e)$ *and with*

$$I = \frac{1}{\mathbb{E}[\tau_\epsilon]}\mathbb{E}\Big[\sum_{r=1}^{\tau_\epsilon} J_r\Big], \qquad D = \frac{1}{\mathbb{E}[T-\tau_\epsilon]}\mathbb{E}\Big[\sum_{r=\tau_\epsilon+1}^{T} D_r\Big],$$

the inequality $\mathbb{E}\big[S_{T\vee\tau^\epsilon}\big] \leq \mathbb{E}[T](1 + V(\epsilon, N))$ *holds, where* $V(\epsilon, N) = \frac{R_i}{I}\big(\sqrt{\epsilon}N\big) + \sqrt{\epsilon}\frac{N}{\mathbb{E}[T]I}$.

Therefore, with (23), we get the desired upper bound by appropriately setting I and D as in the lemma. Hence, we get

$$\frac{-\log P_e}{\mathbb{E}[T]} \leq \frac{D}{1-\Delta}\Big(1 - \frac{R}{I} + U(P_e, M, \epsilon) + V(\epsilon, N)\Big).$$

One can show that for any $(M^{(n)}, N^{(n)})$-VLCs as in Definition 3 the residual terms U, V, Δ converge to zero as $n \to \infty$.

Conclusion

This paper presents an upper bound on the feedback error exponent and feedback capacity of channels with stochastic states, where the states evolve according to a general stochastic process. The results are based on the analysis of the drift of the entropy of the message as a random process.

REFERENCES

[1] C. Shannon, "The zero error capacity of a noisy channel," *IEEE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, sep 1956.

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

[3] R. L. Dobrushin, "An asymptotic bound for the probability error of information transmission through a channel without memory using the feedback," *Probl. Peredachi Inf.*, vol. 8, pp. 160–161, 1962.

[4] E. A. Haroutunian, "Lower bound for error probability in channels with feedback," *Probl. Peredachi Inf.*, vol. 13, pp. 36–44, 1977.

[5] A. Y. Sheverdyaev, "A lower estimate for the probability of error in a discrete memoryless channel with feedback," *Probl. Peredachi Inf.*, vol. 18, no. 4, pp. 5–15, 1982.

[6] A. Goldsmith and P. Varaiya, "Capacity, mutual information, and coding for finite-state markov channels," *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 868–886, may 1996.

[7] G. Como, S. Yuksel, and S. Tatikonda, "The error exponent of variable-length codes over markov channels with feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2139–2160, may 2009.

[8] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 323–349, jan 2009.

[9] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 644–662, feb 2009.

[10] H. Yamamoto and K. Itoh, "Asymptotic performance of a modified schalkwijk-barron scheme for channels with noiseless feedback (corresp.)," *IEEE Transactions on Information Theory*, vol. 25, no. 6, pp. 729–733, nov 1979.

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

[12] B. Nakiboglu and R. G. Gallager, "Error exponents for variable-length block codes with feedback and cost constraints," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 945–963, mar 2008.

[13] R. Ahlswede and N. Cai, "The AVC with noiseless feedback and maximal error probability: a capacity formula with a trichotomy," in *Proceedings of IEEE International Symposium on Information Theory*. IEEE, 1997.

[14] A. Tchamkerten and I. Telatar, "On the universality of burnashev's error exponent," *IEEE Transactions on Information Theory*, vol. 51, no. 8, pp. 2940–2944, aug 2005.

[15] ——, "Variable length coding over an unknown channel," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 2126–2145, may 2006.

[16] J. Chen and T. Berger, "The capacity of finite-state markov channels with feedback," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 780–798, mar 2005.

[17] J. H. Bae and A. Anastasopoulos, "The capacity of markov channels with noiseless output and state feedback," in *2010 Information Theory and Applications Workshop (ITA)*. IEEE, jan 2010.

[18] A. Anastasopoulos and J. Wu, "Variable-length codes for channels with memory and feedback: Error-exponent lower bounds," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, jun 2017.

[19] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Transactions on Information Theory*, vol. 9, no. 3, pp. 136–143, jul 1963.

[20] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback–i: No bandwidth constraint," *IEEE Transactions on Information Theory*, vol. 12, no. 2, pp. 172–182, apr 1966.

[21] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, 1998.

[22] M. Heidari, A. Anastasopoulos, and S. S. Pradhan, "On the reliability function of discrete memoryless multiple-access channel with feedback," in *2018 IEEE Information Theory Workshop (ITW)*. IEEE, nov 2018.

[23] J. L. Massey, "Casuality, feedback and directed information," in *Symposium on Information Theory and its Applications*, 1990.