Rate-Delay Tradeoffs for Communicating a Bursty Source over an Erasure Channel with Feedback

Brian P. Dunn and J. Nicholas Laneman
University of Notre Dame
Dept. of Electrical Engineering
Notre Dame, IN 46556
Email: {bdunn2, jnl}@nd.edu

***Abstract*—We consider the protocol overhead of meeting an expected delay constraint in sending a bursty source over a binary erasure channel with feedback. In contrast to related work on stability and delay for medium-access control, we focus on physical layer aspects of the problem by jointly considering queuing and message encoding for transmission over the channel. Our model necessitates the lossy encoding of source idle times, and we investigate the amount of information implicitly communicated by their efficient encoding. We show that the stable throughput region for large, but finite, delay is the same as without this additional constraint; however, our results suggest that achieving the same average delay requires a significant increase in channel bandwidth. Outer bounds on the achievable rate-delay region are also developed.**

I. Introduction

In novel work by Gallager [1], a lower bound on the amount of protocol information required to meet an expected delay constraint is found. The formulation in [1] is surprisingly general in nature, and makes few assumptions about a particular network topology. However, to establish (achievable) upper bounds on protocol overhead, Gallager relies on an arbitrarily large number of sources and error-free transmissions over the network.

We are motivated by this work to investigate the impact of random message arrivals on bandwidth requirements for point-to-point links, and to study how channel errors affect Gallager's formulation of protocol overhead. Specifically, as in [1], we consider an average delay constraint, but instead of a general noise-free network we consider point-to-point communication over a channel that introduces errors in the transmissions. For simplicity of exposition, in this paper we focus on the case of a binary erasure channel (BEC) with feedback. Our results suggest that if an application requires low average delay, transmission over the BEC with feedback may require more protocol overhead than what is required for an error-free network with the same average delay, such as that considered in [1]. However, the stability region is the same in both cases, indicating that for sufficiently large average delays any additional protocol overhead becomes negligible.

II. Model and Formulation

Consider the communication system depicted in Figure 1. Messages $\mathsf{M}_1, \mathsf{M}_2, \ldots$ arrive at the encoder at times $\mathsf{T}_{s,1}, \mathsf{T}_{s,2}, \ldots$ and are encoded and sent to the destination over a BEC with erasure probability ϵ. Let $\underline{\mathsf{M}}_i = (\mathsf{M}_i, \mathsf{T}_{s,i})$ denote the ith message-arrival time pair. Every $1/R$ seconds the encoder outputs a channel symbol $\mathsf{X} \in \{0, 1\}$, and the decoder receives $\mathsf{Y} \in \{0, 1, e\}$ according to the channel law

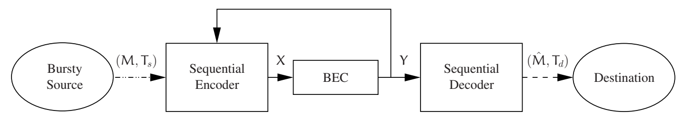

Fig. 1. Block diagram depicting a generic encoder-decoder for the transmission of a bursty source. Messages arrive at the encoder at random time instances, but are revealed to the decoder at discrete periodic time instances corresponding to when the decoder first receives information about the message.

$$\mathsf{Y} = \begin{cases} \mathsf{X} & \text{with probability } 1-\epsilon, \\ e & \text{with probability } \epsilon. \end{cases} \tag{1}$$

Upon transmission of a channel symbol the output of the BEC is immediately available to the decoder and, via a noise-free feedback link, also to the encoder. The corresponding channel capacity is therefore $R(1-\epsilon)$ bits/second, and can be achieved by sequentially retransmission of uniformly distributed bits until each is successfully received.

The decoder uses the sequence of received symbols $\mathsf{Y}_1, \mathsf{Y}_2, \ldots$ to form an estimate $\hat{\mathsf{M}}_1, \hat{\mathsf{M}}_2, \ldots$ of the message sequence. It is assumed that the decoder reconstructs each message symbol as soon as possible, and these decoding times are denoted as $\mathsf{T}_{d,1}, \mathsf{T}_{d,2}, \ldots$. The delay for message M_i is a random variable given by $\mathsf{D}_i = \mathsf{T}_{d,i} - \mathsf{T}_{s,i} > 0$, and the expected delay averaged over all messages is defined as

$$D := \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}[\mathsf{D}_i], \tag{2}$$

when the limit exists.

In order to state our main results, we require some definitions.

A *bursty source* $\mathcal{S}$ is characterized by the following:

- A continuous-time counting process $\mathsf{N}_s(t)$ specifying the number of messages generated up to and including time t.

- A message alphabet $\mathcal{M}$ that is finite or countable, and an associated distribution on message symbols $P_{\mathsf{M}}(\mathsf{m})$, $\mathsf{m} \in \mathcal{M}$.

The definition of an encoder-decoder can be made similar to the standard form if we consider the input to the encoder to be a super-symbol that is a collection of all messages received over the last $1/R$ seconds. Message ordering must be preserved at the decoder, so the encoder also needs partial knowledge of the timing information. We compute the delay between message arrivals and their reconstructions, so it is natural to include exact timing information in the 'super-symbols.' This also gives the encoder sufficient information to determine message orderings.

A *super-symbol* U *of rate* R *for the source* $\mathcal{S}$ is a collection of messages and their corresponding arrival times over a duration of $1/R$ seconds corresponding to the time interval between two consecutive channel uses. Note that if there are no message arrivals during a given time interval, then the corresponding super-symbol is $\mathsf{U} = \varnothing$. If the alphabet for a message-arrival time pair is $\underline{\mathcal{M}} = \mathcal{M} \times \mathbb{R}^{+}$, then the alphabet for U is $\mathcal{U} = \varnothing \cup \underline{\mathcal{M}}^{*}$, where the notation $\mathcal{X}^{*}$ is defined as $\mathcal{X}^{*} := \bigcup_{k=1}^{\infty} \mathcal{X}^{k}$.

A generic sequential encoder for the BEC with feedback outputs a channel symbol at each time interval that is a function of all previous message super-symbols and all previous received channel symbols. The corresponding generic sequential decoder outputs message reconstructions as soon as possible, but has access to the entire history of received channel symbols each time it forms a reconstruction.

Formally, a *sequential encoder-decoder* (F, G, R) *for the sequence of super-symbols* $\mathsf{U}_1, \mathsf{U}_2, \ldots$ *of rate* R consists of

- A sequence of encoder mappings

$$F_i : \mathcal{U}^{i} \times \mathcal{Y}^{i-1} \rightarrow \{0, 1\}, \quad i = 1, 2, \ldots \tag{3}$$

- A sequence of decoder mappings

$$G_i : \{0, 1, \epsilon\}^{i} \rightarrow \mathcal{U}, \quad i = 1, 2, \ldots \tag{4}$$

Note that in (3) and (4) i is a time index, not the block-length. Again, the output of the encoder at time $t = i/R$ is a function of all messages that arrived up to that time and all previously received channel symbols, i.e., $\mathsf{X}_i = F_i(\mathsf{U}_1, \mathsf{U}_2, \ldots, \mathsf{U}_i, \mathsf{Y}_1, \mathsf{Y}_2, \ldots, \mathsf{Y}_{i-1})$, and the output of the decoder at time $t = i/R$ is a function of all channel symbols received up to that time, i.e., $\tilde{\mathsf{U}}_i = G_i(\mathsf{Y}_1, \mathsf{Y}_2, \ldots, \mathsf{Y}_i)$. The definitions allow the encoder and decoder's memory to go to infinity, but they do not concern delay. Note that the output of G_i is *not* a reconstruction of U_i specifically, but is an arbitrary super-symbol containing a set of message reconstructions. Furthermore, the set of message reconstructions contained in $\tilde{\mathsf{U}}_i$ need not be the same set as that contained in *any* particular U_i.

The key difference between the definition of a sequential encoder-decoder given here and the standard information theoretic encoder-decoder definition is that here encoders and decoders are indexed by time but not by *message* arrivals. One purpose of defining super-symbols is to allow the encoding and decoding functions to be indexed by time and *super-symbol* arrivals. This problem does not arise in the standard information theoretic context because time indexing and message arrival indexing are equivalent. Note that the latter scenario corresponds to a 'backlogged' queue, where the encoder always has messages to send. However, if the queue is not backlogged the physical layer does not automatically differentiate between meaningful packet data and source idle times when the encoder has no useful (message) information to send.

Definition 1: A rate-delay pair (R, D) is said to be *achievable* if there exists an (F, G, R) sequential encoder-decoder for which the limit

$$\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}[\mathsf{D}_i] \leq D \tag{5}$$

exists, and $\hat{\mathsf{M}}_k = \mathsf{M}_k$ with probability 1, for $k = 1, 2, \ldots$. The *rate-delay region* for a source-channel pair is the closure of the set of achievable rate-delay pairs (R, D).

Definition 2: A rate R is said to be *achievable* if for some $D < \infty$ the rate-delay pair (R, D) is achievable. The *stability region* for a source-channel pair is the closure of the set of achievable rates.

Throughout the remainder of the paper, we focus on a particular source model that highlights the issue of burstiness while keeping the analysis tractable. Specifically, we assume the source $\mathcal{S}$ satisfies:

- Messages are memoryless uniform bits, i.e., $\mathsf{M}_k \in \{0, 1\}$ are i.i.d. with $P_{\mathsf{M}}(0) = P_{\mathsf{M}}(1) = 1/2$.
- Message arrivals are Poisson, i.e., $\mathsf{N}_s(t)$ is the counting process for a homogeneous Poisson process with rate parameter λ. Therefore, the number of messages in any two disjoint intervals of $1/R$ seconds, $\mathsf{N}(t) = \mathsf{N}_s(t) - \mathsf{N}_s(t - 1/R)$, are independent Poisson random variables with parameter λ.

III. Bounds on the Achievable Rate-Delay Region

In this section we present our main results. First, we give two independent outer bounds on the achievable rate-delay region. The bounds are intriguingly different in nature, but each is active in some rate/delay regime for any value of ϵ. Second, we give an inner bound on the achievable rate-delay region by considering a family of simple protocols that are practically motivated but fail to meet the outer bounds. Finally, we show that as intuition suggests, the stability region is $\{\lambda : \lambda < R(1 - \epsilon)\}$.

A. Outer Bounds

1) Gallager Bound: Consider the mutual information between N message-arrival time pairs at the encoder and their corresponding reconstructions at the decoder given by $\mathbb{I}(\underline{\mathsf{M}}^{N}; \hat{\underline{\mathsf{M}}}^{N}) = \mathbb{I}(\mathsf{M}^{N}, \mathsf{T}_s^{N}; \hat{\mathsf{M}}^{N}, \mathsf{T}_d^{N})$. Because messages and arrival times are mutually independent, we have

$$\mathbb{I}(\mathsf{M}^{N}, \mathsf{T}_s^{N}; \hat{\mathsf{M}}^{N}, \mathsf{T}_d^{N}) \geq \mathbb{I}(\mathsf{M}^{N}; \hat{\mathsf{M}}^{N}) + \mathbb{I}(\mathsf{T}_s^{N}; \mathsf{T}_d^{N}). \tag{6}$$

The channel can support at most $1-\epsilon$ bits per channel use (on average). Following [1], the number of channel uses per second R required for reliable communication over the BEC is therefore bounded by

$$R \geq \frac{1}{1-\epsilon} \liminf_{N\to\infty} \frac{1}{N}\left[\mathbb{I}(\mathsf{M}^N;\hat{\mathsf{M}}^N) + \inf_{\mathcal{P}_N} \mathbb{I}(\mathsf{T}_s^N;\mathsf{T}_d^N)\right], \quad (7)$$

where the infimum is over the set of joint distributions $\mathcal{P}_N$ satisfying the average delay constraint (2), as well as the relevant conditions in [1] on the marginal distributions. In (7), $\mathbb{I}(\mathsf{M}^N;\hat{\mathsf{M}}^N)/N = \lambda$, the arrival rate of the source λ times the source entropy, which here is 1 bit per source symbol. As elucidated in [1], $\mathbb{I}(\mathsf{T}_s^N;\mathsf{T}_d^N)$ represents information about message arrival times that is *somehow* being sent to the decoder, regardless of whether or not it is intended to be conveyed. The associated rate-distortion problem provides a lower bound on the protocol overhead required to meet the delay constraint—for any protocol or transmission strategy. The main result of [1] is a lower bound on protocol overhead, which leads to the following outer bound on the achievable rate-delay region:

$$R \geq \frac{1}{1-\epsilon}\left[\lambda - \log_2(1-e^{-\lambda D})\right]. \quad (8)$$

2) Genie-aided Bound: Consider an encoder that retransmits messages as they arrive, uncoded, until each is successfully received, and sends an arbitrary channel symbol when there are no messages to send. In our model, the decoder would have no way of determining which successfully received channel symbols correspond to valid messages and which do not contain any information about the message sequence. However, if the decoder has access to side information indicating the position of valid messages within the sequence of received channel symbols, it can losslessly reconstruct each message. Furthermore, symbol retransmission achieves capacity for the BEC with feedback while minimizing average delay. Accordingly, the genie-aided encoder-decoder does at least as well as any sequential encoder-decoder, so the genie-aided rate-delay region outer bounds the (non-aided) rate-delay region.

Observe that the genie-aided encoder's message queue can be modeled as an M/G/1 queue with Poisson arrivals of rate λ and geometrically distributed service times with mean service rate $R(1-\epsilon)/\epsilon$. Because of the slotted channel model, unlike for an M/G/1, a message arriving to an empty queue must wait until the next channel slot to begin transmission (service). In fact, regardless of the queue length, because of the slotted channel, each message waits an average of $1/2R$ seconds before entering the encoder's queue. Therefore, the expected delay for the genie-aided encoder-decoder is the sum of the average slotted delay and the expected time spent passing through an M/G/1 queue. Using the Pollaczek-Khinchin (P-K) formula [2] to evaluate the latter expectation, we arrive at the following bound on the achievable rate-delay region:

$$D \geq \frac{\lambda(1+\epsilon)}{2R(1-\epsilon)\left[R(1-\epsilon)-\lambda\right]} + \frac{\epsilon}{R(1-\epsilon)} + \frac{1}{2R}. \quad (9)$$

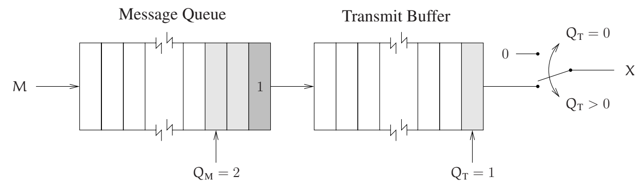

Fig. 2. Block diagram for fixed source sample-variable rate encoding (FV), depicting the Message Queue cascaded with the Transmit Buffer. The number of messages in the Message Queue (besides the start bit of '1' in the darker box) is denoted $\mathrm{Q_M}$, and the number of symbols in the Transmit Buffer (including any start bits present) is $\mathrm{Q_T}$.

3) Outer Bound Comparison: The P-K formula gives the expected throughput-delay tradeoff for a packetized model. However, it fails to capture the transfer of protocol information beyond what is contained in messages. For example, the encoder must somehow inform the decoder of when a valid message is being sent over the channel. This can be seen by observing that for certain average delays (or correspondingly channel uses per second), the Gallager bound is strictly tighter than the P-K bound. For simplicity, consider $\epsilon = 1$ and $\lambda = 1$. The P-K bound in (9) simplifies to $D \geq 1/[2(R-1)]$ and (8) becomes $R \geq 1 - \log_2(1-e^{-D})$. By combining both bounds we see that for values of average delay satisfying $(1-e^{-D})^D > 1/\sqrt{2}$, i.e., $D \in (0.206, 1.68)$, the Gallager bound is strictly tighter than the P-K bound. This can be seen graphically in Figure 3.

B. Inner Bounds

We now compute inner bounds on the achievable rate-delay region by considering a particular coding strategy that maps a fixed number of messages N to a variable number of channel symbols B. The encoder waits to accumulate N messages, then sends a start bit of '1' to indicate the beginning of a transmission followed by each of the N messages sequentially. A block diagram for the fixed source sample-variable rate encoder (FV) is depicted in Figure 2, showing the encoder's architecture as a Message Queue cascaded with a Transmit Buffer. The number of messages in the Message Queue (besides the start bit of '1' in the darker box) is denoted as $\mathrm{Q_M}$, and the number of symbols in the transmit buffer is $\mathrm{Q_T}$. The block size N is a design parameter that is chosen in advance and known to both the encoder and decoder. The encoding operation of the FV coder of block size N is described as follows:

1) As messages arrive they enter the Message Queue sequentially.
2) If $\mathrm{Q_M} \geq N$, the first N messages in the Message Queue, in addition to the start bit of '1', are transferred to the Transmit Buffer.
3) Every $t = i/R$ seconds, the encoder must output a channel symbol. If the Transmit Buffer is empty ($\mathrm{Q_T} = 0$), it outputs a '0', otherwise it continuously retransmits the rightmost symbol in the Transmit Buffer until it is successfully received by the decoder. Upon

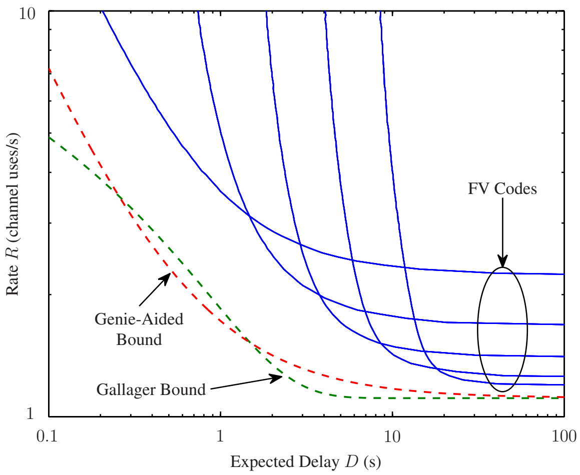

Fig. 3. Outer bounds (dashed) on the rate-delay tradeoff and achievable regions (solid) for $\lambda = 1$ and an erasure probability of $\epsilon = 0.1$. The 'Gallager Bound' (– – –) is the entropy rate of the source plus the lower bound on protocol information given in [1]. The 'Genie-Aided Bound' (– – –) is derived from the expected delay experienced by a message passing through an M/G/1 queue. The achievable rate-delay curves are numerical simulations of FV codes (—) where, for successively lower curves at $D = 100$, $N = 1, 2, 4, 8, 16$, respectively.

successful transmission, the bit is removed from the Transmit Buffer.

The corresponding decoder operates as follows:

1) Wait until a start bit of '1' is successfully received.
2) Declare the next N channel symbols that are successfully received to be the next N message estimates.
3) Goto step 1.

Note that as long as $\mathrm{Q_M}$ and $\mathrm{Q_T}$ remain bounded, the decoder will be able to perfectly reconstruct the message sequence in some finite amount of time. We examine the performance of FV coders for fixed N, in addition to the infimum over all N, i.e., over the class of all FV coders for $N = 1, 2, \ldots$.

If the encoder always has messages to send (is backlogged), the block size N determines the amount of protocol overhead per message. N can be chosen large in order to minimize this overhead, but in doing so the encoder must wait longer to accumulate a sufficient number of messages, thereby increasing message delays. Consequently, there is a tradeoff between the required data rate and the average delay experienced by messages. The performance of FV codes for several block sizes are shown in Figure 3, along with the 'Gallager' and 'Genie-Aided' outer bounds.

C. Stability Region

Heuristically, as $R \to \lambda^+$ the queue lengths become large, and the Transmit Buffer is rarely empty. In this case $\mathrm{B} \gtrsim N+1$, the encoder is always sending useful information, and

$$\mathbb{E}\left[\frac{\text{FV channel uses}}{\text{source symbol}}\right] \to \frac{N+1}{N}. \tag{10}$$

To approximate D as $R \to \lambda^+$, consider $R^* \lesssim \lambda[(N+1)/N]$ as the average service time of an $M/D/1$ queue, and apply the P-K formula [2] to obtain

$$D \overset{R\to 1^+}{\approx} \frac{N+1}{2(NR-N-1)} + \frac{1}{2NR} + \frac{1}{2R}, \tag{11}$$

where for simplicity we take $\lambda = 1$. For any practical scheme N must be finite, but to compute the closure of achievable rates with FV coding we must consider performance as $N \to \infty$. Taking the limit of (11) as $N \to \infty$ gives an asymptotically achievable R for large N with an average delay of

$$\lim_{N\to\infty} D \overset{R\to 1^+}{\approx} \frac{1}{2(R-1)} + \frac{1}{2R}, \tag{12}$$

as long as the limit exists. Thus it seems that for any $R > \lambda$, N can be chosen sufficiently large such that the expected delay remains bounded and R is (by definition) achievable. This statement is made more formally in the following theorem.

Theorem 1: The closure of the set of achievable rates under FV coding is

$$\mathrm{cl}\,(\mathcal{R}_{\mathrm{FV}}) = \{R : R \geq \lambda/(1-\epsilon)\}. \tag{13}$$

Proof: Due to space limitations, a sketch of the proof is given. First, sufficient conditions for $R \in \mathcal{R}_{\mathrm{FV}}$ are given as a function the average number of channel uses per second required by FV coding to transmit the source, denoted by R^*. Next, an upper bound on R^* is derived by conditioning on the Message Queue being empty at the end of the previous block transmission. Finally, it is shown that as $N \to \infty$ any R satisfying $R > \lambda$ is achievable. ■

IV. Conclusions

In this paper we examined the transmission of a bursty source over the binary erasure channel with feedback subject to an expected delay constraint. Inner and outer bounds on the achievable rate-delay region were obtained, and the stability region was found to be identical to that for the typical packetized model. Our results suggest that although the stability region is not reduced, there is unavoidable overhead associated with the lossy encoding of queue idle times, something that is often overlooked. Furthermore, the implicit transfer of information about queue states through the encoding of idle times may play a key role in understanding the overhead associated with channel access.

The protocols used in this paper are inherently limited to the BEC with feedback, and therefore inner bounds for more general noisy channels will require the pursuit of alternative techniques. However, our results illustrate that even for the simplistic BEC with feedback, satisfying a delay constraint over a noisy channel may require additional protocol overhead beyond what is necessary in a noise-free network.

References

[1] R. G. Gallager, "Basic Limits on Protocol Information in Data Communication Networks," *IEEE Trans. Inform. Theory*, vol. 22, no. 4, pp. 385–398, July 1976.

[2] D. P. Bertsekas and R. G. Gallager, *Data Networks*, Second ed. Prentice Hall, 1991.